Learning to Identify New Information

Barry Schiffman

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2005

ABSTRACT

Learning to Identify New Information

Barry Schiffman

This thesis is an investigation into a new problem in natural language processing: new-information detection. It is a similar task to first-story detection, but with a very large difference. First-story detection operates on the document level, while new-information detection is on the statement level. In its fundamental guise, new-information detection is the ability for a machine to be able to compare two textual statements and decide whether they say the same thing or not. But the task is complicated by the fact that each new statement must be tested against all previous statements.

In this thesis, I show that the sentence is a poor choice of syntactic unit for this task since sentences are arbitrarily composed of one or more structures. Thus, the system must do a deeper syntactic analysis of the inputs than recognizing sentence boundaries. At the same time, I found that context is important, and I developed a mechanism to look beyond sentence boundaries for evidence of novelty. Thus, the system I developed considers a mixture of features, from a micro perspective, looking within sentences, and from a macro perspective, looking beyond sentence boundaries. I apply machine learning techniques to combine the features coherently into a unified hypothesis for the problem, using rule induction. The system is designed to function in a multi-document summarization system, like Columbia's NewsBlaster, for which it produces update summaries focusing on the developments of the day in an event that has interested the public over a period of several days. The new-information system provides all the novel statements to the DEMS summarizer, which I had previously built for NewsBlaster, for the final selection of material.

The system also includes a semantic unit that improves performance a bit, but not as much as I hoped. The system uses a plugin lexicon that is largely taken from the WordNet data at present.

Contents

List of Figures

List of Tables

Acknowledgments

More than anyone, my adviser, Kathy McKeown, deserves my gratitude for giving me the opportunity to have undertaken this wonderful adventure and challenge. By all rights, I was an unlikely computer science grad student, but Kathy is nothing if not courageous. And over the years, her encouragement, guidance and support have been crucial. It was always Kathy who knew what we had to do next and who kept the work moving forward. I also thank the others on my committee for their patience and tolerance, especially Inderjeet Mani for our collaboration and friendship.

My list of creditors at this great university is long. From my officemates and colleagues in the Natural Language Processing Group, I learned countless things about our craft and about the world. I so appreciate their kindness and generosity. Nor do I want to leave out the faculty, staff and students of the department whose dedication to excellence has been awesome.

Beyond the university, there are so many people to thank. I often think of two former colleagues at The New York Times who told me as I was about to leave that they envied me for having found a passion. They made clear that I was doing the right thing. I wish I could list all of my many other friends for their encouragement and support over the years as I reinvented myself. But they will have to remain anonymous.

I only wish my parents were alive to relish this moment. They were working people who came to this country when they were young to give their children the opportunities that do not exist anywhere else. But I have my daughters, Jennifer, Maryanne and Elizabeth, whose affection is always close to my heart. And Zina Saunders, of course, my partner and constant companion. She made the whole process possible in so many ways. She continues to amaze me with her generosity and kindness and wonderful optimism, assuring me day in and day out that I could really do this.

To Zina: this is my drawing table

Chapter 1

Introduction

This dissertation proposes an automatic means of monitoring a topical stream of text, filtering out information that has already been seen and presenting the user with only new material. In this task, new information detection, information is considered in small, discrete units – clauses. As such, one unit is a single standalone statement about some event or condition in the world.

I have built a system, the New Information Agent, or **NIA**, that operates on topical clusters of news documents. The program reads each article in turn and selects only new material – the atomic facts in the current article that are not covered by the accumulated facts from earlier articles. This system can be used as a summarizer that organizes its output on the principle that the newest information available is the most important information. In effect, it is a kind of bulletin service that raises an alert when something new has been received and provides an update summary. There are already several news browsing or aggregation services that automatically collect and cluster news from many different sources, Google for example (`http://www.google.com`). Columbia's NewsBlaster[MBE+02] not only collects and clusters the news, but also provides automatic summarization of the clusters (`http://newsblaster.cs.columbia.edu`) as does the News in Essence system at the University of Michigan (`http:www.newsinessence.com`). The World Wide Web editions of many news organizations like the New York Times (`http://www.nytimes.com`) or the

Washington Post (`http://www.washingtonpost.com`) offer readers headline or alert services in which they email new articles containing keywords selected by the readers. The goal in developing **NIA** is to improve on this service; it can filter out the repetitive material from the new articles and emphasize what has changed in the world.

New information detection is related to a broader inquiry known as novelty detection, where patterns are used to search for anomalies in diverse fields, including signal processing, computer vision, pattern recognition, data mining and robotics. In information retrieval, first story detection is a novelty detection problem, where the anomaly is the first story that does not fit into an existing topic cluster. First-story detection is distinct from my goal in new information detection. Where first story detection operates at the document level, **NIA** operates at the level of single statements. First story detection relates to topic tracking and clustering in that it decides whether the next document continues an old topic or whether it begins a new one. The work in this thesis concerns a fine-grained examination of short spans of texts, looking at sentences and subsentence chunks to determine whether a particular fact was already seen or not.

In broad terms, the goal in new information detection is an ambitious Artificial Intelligence undertaking to emulate the behavior of humans performing sophisticated tasks. While the software presented here operates on news, there is no reason it could not be modified to serve the needs of a physician who needs to read numerous medical journals to keep up with new developments in his or her specialty, or a lawyer who is searching through many judicial decisions for new precedents, or a financial analyst who needs to view disparate kinds of material for events that could move the markets, or a government analyst struggling to discern political shifts around the world.

The central issue in this thesis is to develop how a computer program that, given some number of *new documents* and some store of previously scanned information, can detect what exactly is new in the input documents, add it to the store of knowledge and return to the user a precis of the new material. Given the large amount of reading that many professionals must do, such a program would be extremely valuable.

The inputs to the system have to be processed so that the system receives clusters of documents on one topic. The complexity of the preprocessing could vary greatly depending

on the task. In NewsBlaster, news sites are crawled automatically and collected without any topical cues. NewsBlaster uses a clustering algorithm to divide the full collection into coherent groups of 5 to 35 articles. In some domains, the clustering might be improved by some meta information, in others, clustering might be even more difficult, but that is not the focus of this research.

After the clusters are processed and the system returns the new information – a set of the original statements that are classified as new – the output could be further filtered according to the users' needs. The system I describe, operating on news, can be fed through a multi-document summarizer, such as the one I developed, DEMS [SNM02], Dissimiliarity Engine for Multi-document Summarization, which seeks out the passages with the most *news value*.

The task performed by **NIA** is the focus of the dissertation: what kind of linguistic features are necessary to discover novelty, what sort of rules can use these features to make the novelty decision, and what learning method will best produce these rules. The results are embodied in the working system, **NIA**.

NIA needs functionality that is closer to text understanding than many real-world natural language processing (NLP) systems offer. Summarization systems relying on statistical distributions of words do not require semantics. The best search engines, which are of enormous value for millions of people, can search billions of documents at a speed many magnitudes faster than humans, but they can fail if even a meager inference must be drawn from the query. Yet, this dissertation does not claim to achieve text understanding, which is likely to remain out of reach for some time to come, but, instead, it determines what surface tools and techniques can be utilized to solve the problem at hand.

1.1 New Problem

New information detection, is a new area of inquiry in both NLP and IR, and naturally arose from recent advances in both communities. In NLP, researchers have moved from summarization of single documents to summarization of sets of documents, creating the need for effective ways to identify repetition in the input documents and avoid repetition in

the summary. In IR, researchers are focusing on various ways of retrieving and combining smaller units of text than whole documents. They are exploring ways to match narrow queries to passages within documents. Even question answering is similar since questions are matched to answers. Again, these tasks would be greatly enhanced by software to distinguish new material from that which repeats the notions realized in previously selected passages.

1.1.1 DUC

At the recent Document Understanding Conferences (DUC), researchers were asked to make relatively short summaries of ten or so documents closely clustered around a topic or theme. Since the beginning of the conferences, NIST, the organizer, has asked participants to produce general summaries from clusters where the themes were not identified or linked to a question or query. The clusters themselves implicitly defined a theme. At the last two conferences, NIST added tasks that, in one year, asked participants to focus on a query or topic, and, in another, to produce a biographical summary on a given individual. Duplication hurts the systems in two ways: First, it detracts from fluency because the same information is stated in different ways, and second, it reduces coverage by wasting space on the unnecessary repetition. At the most recent DUC, the results on the quality questions[1] showed that duplication was a serious problem even in these very short summaries, 100 words maximum. Figure 1.1 shows that human judges found some duplication problem in nearly 40% of the machine summaries. In contrast, only 11% of the human-written summaries had any duplication, and none of those cases were serious.

The scoring was on a five point scale with "1" indicating no problem, and "5" indicating "quite a lot." If a penalty scale is imposed to assign 1 point of duplication for a "2" rating, 2 points for a "3" raging, and so forth, the expected penalty value for humans would be 0.15, but 0.68 for machines.

[1]DUC results were scored in two ways at the 2004 conference. An automatic system, Rouge, produced scores based on n-gram overlaps and longest common subsequences. Manual scores covered both the content of the entries and their readability. The quality questions in general concerned readability, andthey included duplication.

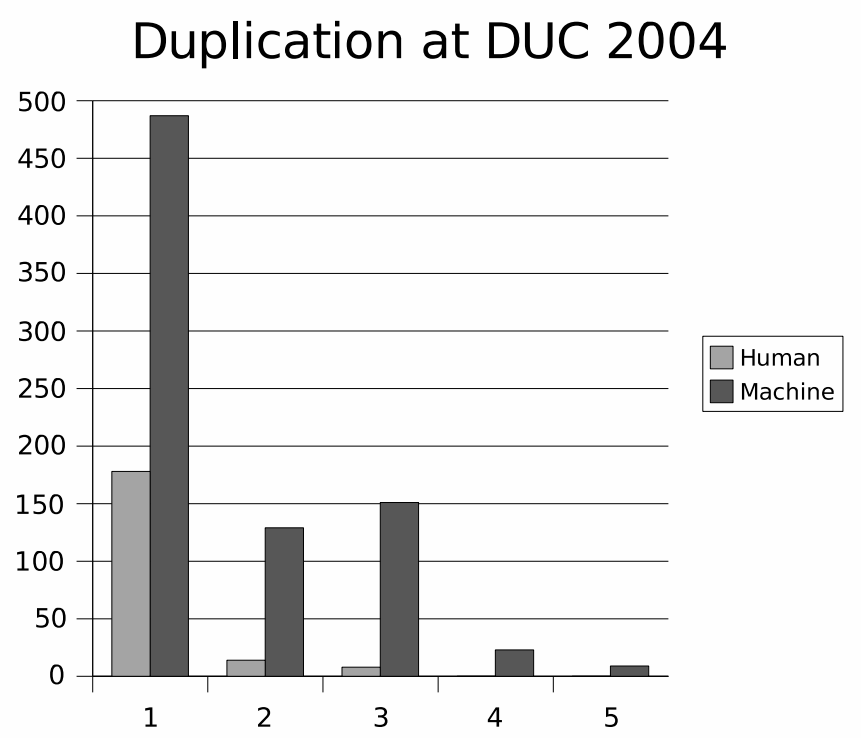

Figure 1.1: Duplication is a problem even in making short generic summaries of 100 words each. This charts shows the experience at DUC 2004. The x-axis identifies the categories, and the y-axis shows the number of summaries in each category. The duplication question asked of the humans who reviewed these summaries was, "To what degree does the summary say the same thing over again." A rating of 1 means "None"; 2, "Minor"; 3, "some"; 4, "more than half"; 5, "quite a lot". There were 200 human written summaries and 799 machine produced summaries. The first column shows the number of summaries without duplication problems. Thus 178, or 89%, of the 200 human summaries had no duplication, but only 479, or 60%, of the 799 machine summaries were free of repetition. Even more telling is the fact that only 11% of the human summaries had either "minor" or "some" duplication, while 35% of the machine summaries were in those categories. No human summaries had more serious problems, but 2% of the machine summaries did.

Over the years, DUC summary lengths have ranged from 10 words to 400 words for each cluster. With such stringent length restrictions, duplication can be a costly flaw for a participating system. It's easy to see just how clumsy a summary can appear when it has repetitive information. Below is one of the summaries submitted last year:

> Duisenberg, who spoke at an event in London, said the Dec. 1 and Dec. 22 meetings of the Central Bank's policy-making body will gauge the outlook for inflation and the EU economy.
> In a surprise move, nations adopting the new European currency, the euro, **dropped key interest rates** Thursday, effectively setting the rate that will be adopted throughout the euro zone on Jan. 1.
> Ten of the 11 countries adopting the euro **dropped their interest rate** to 3 percent.
> Making their first collective decision about monetary policy, the 11 European nations launching a common currency on Jan. 1 **cut interest rates** Thursday in a surprise move that won market confidence.

This summary[2] captures only one of the key points that the four human-written models cover: that interest rates were cut by European Union members, but it repeats that information three times. The assessors gave this a rating of 5, which indicates "quite a lot" of repetition (See Appendix A).

1.1.2 TREC

At the last three Text Retrieval Conferences (TREC), a new *Novelty Track* was offered, combining a new passage-retrieval task with a novelty task – which was cast as the removal of duplicate information at the sentence level. I took part in the tasks that isolated the novelty problem alone to help in the development of my system.

The inputs for the participants were topics – brief descriptions of an issue or event – and clusters of 25 relevant documents. Each topic provides a title, a description and a narrative, like the example below[3]:

Title India and Pakistan Nuclear Tests

[2]The summary was submitted by the Fudan University Group for cluster D30033.

[3]This was Topic 55 for the 2004 evaluation; the corpus consisted of 14 Associated Press articles and 11 Xinhua articles.

Description On May 11 and 13, 1998 India conducted five nuclear tests; Pakistan responded by detonating six nuclear tests on May 28 and 30th. This nuclear testing was condemned by the international community.

Narrative Relevant documents (*i.e. passages*[4]) mention the nuclear testing conducted in May 1998 by both India and Pakistan. Historical information about the antagonism and rivalry between the two countries is not relevant. Mention of the furor created around the world by these detonations is relevant.

The topics can be considered hypothetical queries over the World Wide Web, a private text collection or documents or text database.

For the retrieval part, the systems were required to identify as many of the sentences that are relevant to the topic as they could find.

In the first year, the output of this part, an ordered list of sentences, was required to be used as the input to the novelty part, in which systems were to remove sentences that duplicated any previous sentence. The results showed that the retrieval task overwhelmed the novelty task, which was entirely dependent on the quality of the retrieval system. In addition, more than 90% of the sentences that the human judges picked as relevant were also determined to be novel. After TREC 2002, experiments by the author and by others using the judges' lists of novel sentences showed that it was very difficult to improve – in terms of the standard metric, the F-measure – on a strategy of accepting all the sentences as novel.

In the second year, NIST increased the amount of duplication within the clusters of articles so that only about 65% of the relevant sentences were novel, and gave systems the option of performing the novelty part alone, independent of the relevance part. After participants submitted the results for the relevance parts, the human selections of relevant material were made available for experiments on the true list of relevant sentences.

The topic above is a typical one. In all there were 536 sentences in the cluster of relevant documents, but only 56 were marked as relevant to the topic. Of these 56 relevant sentences,

[4]The topics were originally used for other TREC evaluations that were about searching for documents. The Novelty reused many of these topics.

21 were marked as novel by the assessor.

For a closeup look at the task, consider a pair of passages. The first is marked as relevant and *novel* sentence from one document in the topic cluster. And the second, composed of two sentences marked as relevant from a later document:

A Some countries or international organizations expressed concerns after Pakistan exploded five nuclear devices on Thursday in response to the nuclear tests carried out by India on May 11 and 13.

B Pakistani Prime Minister Nawaz Sharif announced that his country has successfully conducted five nuclear tests.

Pakistan's move followed India's nuclear testing two weeks ago which triggered worldwide condemnation.

These passages illustrate the deceptive difficulty of the new information detection problem. A human reader would have no trouble to see that the single sentence in the **A** passage obviates the need for both sentences in the **B** passage, but a pure bag-of-words approach – such as comparing the strings of one sentence to those of another lacks the kind of syntactic and lexical knowledge that makes detection of new information possible.

1.1.3 Definition for Evaluation

In both contexts, DUC summarization and the TREC Novelty Track, the problem is set out in a straightforward way, i.e. determining whether there is duplicate content in the system's output. The task sounds simple enough: The DUC Quality Questions say, "To what degree does the summary say the same thing over again?" On its Novelty Track web page, NIST writes, Novel sentences "provide new information that has not been found in any previously picked sentences."

In this work, I will focus on this basic problem, and leave for future work, more philosophical questions of what is *new* or *novel*. In addition, I will postpone efforts to address ways to handle different users' prior knowledge and beliefs about the subject and to deal with the human readers' power to draw inferences. Given some sequence of related events, some people will expect $event_i$ to be naturally followed by $event_{i+1}$ and therefore will not

consider $event_{i+1}$ to be novel; but others will not infer $event_{i+1}$ from $event_i$ and will consider the explicit mention of $event_{i+1}$ to be novel. Finally, I will also set aside the question of estimating the relative importance of novel statements, but it is necessary first to break down the overall problem into smaller, more manageable parts.

This research will treat novelty as an objective quality that will not depend on a user's knowledge or bias. The main focus here will be on seeking a method for making concrete judgments on each of the input statements, a method that will be domain and user independent. Thus I will try to be literal minded in approaching the problem.

In isolation, novelty can be considered the reciprocal of similarity. Two statements either realize the same content or they do not. In practice, the statements do not exist in a vacuum, nor are they limited to expressing a single fact. This means that as each sentence is scanned by the system, each fact in the sentence must be compared against *every fact that came before*. In terms of complexity, the process I am proposing would be $O(n^2)$ like a pairwise comparison of sentences, but there are important differences: The units, clauses in this case, are much harder to extract, requiring parsing, a technology that is still improving; there are more of them, often two or three or more in a sentence. And, complete reference resolution is more crucial with these shorts spans of text. I will discuss my efforts in this area in Chapter 4

To view the difficulty from a different angle, suppose one considered novelty as the reciprocal of similarity. The system is forced to make judgments for each unit in the current document in comparison with all previous units. Thus, a similarity metric needs to be robust enough to find *all* similar statements in previously scanned material. Say statement S_n is being considered. One would apply some metric to each statement that came before, S_1 through S_{n-1}, and eliminate any S_n that scores above a threshold, accepting the rest. In order for the reciprocal to be useful as a novelty detector, the metric has to be a function with a high degree of confidence, but this is not likely. Simfinder, a state-of-the-art similarity tool [HKH+01], operating over paragraphs achieved a precision of 49.3% and recall of 52.9%.

The next section will discuss the application of this research to an existing news browsing system, and the subsequent section will give an overview of my working hypothesis. The

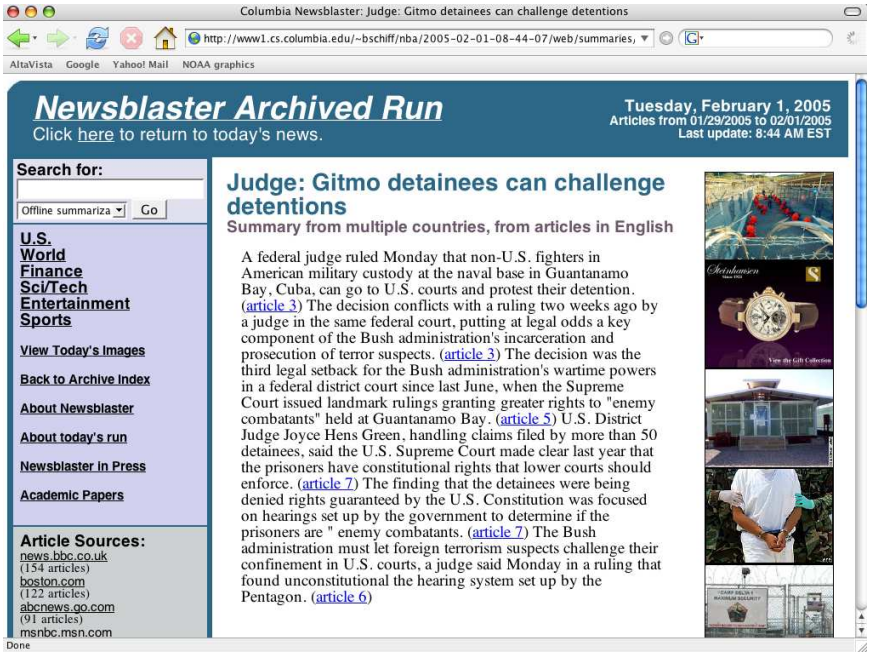

Figure 1.2: A recent summary in Newsblaster on the legal issues involving the detainees at the Guantanamo Bay navy base.

chapter will conclude with the contributions of this work.

1.2 Newsblaster

The results of this research will supply additional functionality to NewsBlaster, an online news browsing system that automatically crawls a large number of news sites on the World Wide Web, gathers the news pages, extracts the texts of the articles, clusters them and prepares a summary of each cluster. The experimental software presented here has been adapted for NewsBlaster to provide a specialized summarization system to highlight developments over time, and operates in conjunction with software to track news events across several days. Clusters of articles in NewsBlaster on a given day range from 4 articles to 40. The web crawling and clustering programs are run once a day at present.

Figure 1.2 shows the Newsblaster interface, showing an archived summary of the news about a decision about the status of the suspected terrorists held at Guantanamo Bay,

Cuba. Along the left underneath the broad categories are a list of the day's sources of news. Along the right side are photos that have been automatically selected to go with the cluster being summarized. The summary was created from the eight articles published on Monday, Jan. 31 (See Figure 1.3), and Tuesday, Feb. 1 (See Figure 1.4), of this year.

This summary of about 200 words was written by DEMS [SNM02], a standalone multi-document summarizer that I built while exploring some strategies for **NIA**. DEMS, the Dissimilarity Engine for Multi-document Summarization, handles most of NewsBlaster summarization. It is a sentence extraction system that uses a linear combination of 12 features to rank sentences. The motivation for DEMS was to provide a way to cope with input clusters of articles that were too diverse to be handled by strategies that rely on strong similarities in the inputs.

In NewsBlaster, important events tend to be closely covered by many news outlets, which are apt to publish multiple articles on the same day. In the competitive world of journalism, the different organizations try hard to distinguish their own work from others, even when all have equal access to all available facts. The summary shown in Figure 1.2 illustrates the kind of variety. There are eight source articles in this cluster. Five of the articles were published in the afternoon or evening of Jan. 31., and they cover nearly every major point. Figure 1.3 shows the first few lines of each of the articles.

Then three more articles were published early the next morning. Figure 1.4 shows the beginnings of these. It's clear that most of the content is the same across all the articles, with no new developments. One statement that stands out in the later group is the comment in the Boston Globe that the ruling is a blow to the Bush Administration. Although that idea is not prominent in any of the earlier articles, it was said before, in the BBC article. A user could pick any one of these eight articles and be reasonably well informed about the event. But there are small differences, details and nuances, that a user might need to know.

This system ultimately aims providing well-crafted summaries, including text generation techniques, but this thesis focuses exclusively on locating the novel and salient content, and will leave generation issues for future work.

Monday, January 31, 2005

MSNBC U.S. District Judge Joyce Hens Green, handling claims filed by more than 50 detainees, said the U.S. Supreme Court made clear last year that the prisoners have constitutional rights that lower courts should enforce.

BBC A US federal judge in Washington has ruled that special military tribunals being used to try hundreds of detainees at Guantanamo Bay in Cuba are illegal.
Judge Joyce Hens Green said the tribunals denied the detainees their basic rights under the US constitution.

CBS A federal judge ruled Monday that foreign terror suspects held in Cuba can challenge their confinement in U.S. courts and she criticized the Bush administration for holding hundreds of people without legal rights. Judge Joyce Hens Green, handling claims filed by about 50 detainees at the U.S. Navy base at Guantanamo Bay, said the Supreme Court made clear last year that they have constitutional rights that lower courts should enforce.

Dallas News The Bush administration must let foreign terrorism suspects challenge their confinement in U.S. courts, a judge said Monday in a ruling that found unconstitutional the hearing system set up by the Pentagon.
U.S. District Judge Joyce Hens Green also raised concerns about whether detainees have been tortured during interrogations. Judges, she said, should make sure people are not detained indefinitely based on coerced and unreliable information.

USA Today WASHINGTON (AP) – A federal judge ruled Monday that some foreign terror suspects held in Cuba can challenge their confinement in U.S. courts and she criticized the Bush administration for holding hundreds of people without legal rights.

Figure 1.3: The beginning paragraphs of the early articles on the court ruling granting Guantanamo detainees the right to go to court.

Tuesday, February 1, 2005

Boston Globe A federal judge ruled yesterday that Guantanamo Bay detainees must be allowed to challenge their imprisonment in court with a defense lawyer, delivering a major blow to the Bush administration's bid to hold suspected fighters in the war on terrorism without judicial oversight.

CNN A federal judge ruled Monday that non-U.S. fighters in American military custody at the naval base in Guantanamo Bay, Cuba, can go to U.S. courts and protest their detention.

The decision conflicts with a ruling two weeks ago by a judge in the same federal court, putting at legal odds a key component of the Bush administration's incarceration and prosecution of terror suspects.

Washington Post A federal judge ruled yesterday that the Bush administration must allow prisoners at the military prison at Guantanamo Bay, Cuba, to contest their detention in U.S. courts, concluding that special military reviews established by the Pentagon as an alternative are illegal. U.S. District Judge Joyce Hens Green said that the approximately 550 men held as "enemy combatants" are entitled to the advice of lawyers and to confront the evidence against them in those proceedings. But, she found, the Defense Department has largely denied them these "most basic fundamental rights" during the reviews conducted at Guantanamo Bay, in the name of protecting the United States from terrorism.

Figure 1.4: The later group of articles.

1.3 Hypothesis

The central hypothesis in this thesis states that it is necessary to look both within the sentence and outside the sentence in order to detect novelty. We identify these two perspectives as the *Micro View* and the *Macro View*, respectively. Each *view* generates several features for each textual unit in the input set. The units used in this dissertation are smaller than sentences; they are intended to express the atomic facts that were introduced above. I use clauses as the units. They are spans of texts that contain a verb and all its arguments, including ellipses. From each clause, a number of features are computed to provide useful as evidence of novelty. For example, the first mention of a person's name is a feature that would clearly be evidence of novelty.

The input texts are analyzed both syntactically and semantically. The syntactic analysis is accomplished with a probabilistic parser to locate the clause boundaries, and the semantic analysis uses a named-entity recognizer and dictionary to equate different surface realizations of references to the same underlying entity. Once this is done, the program computes the features for each unit, i.e. each clause. If the system is run in learning mode, it applies a machine learning algorithm to the feature vectors representing each unit and constructs a set of rules for classifying units. If it is run in classification mode, it applies previously learned rules to the units.

Figure 1.5 shows why a finer-grained approach is both necessary and feasible. Sentences are composed of one or more clauses. The topics of each component are obviously related to one another in some way, but there are often many words and phrases that should not be linked. The full sentence in the Figure reads:

> **Her ruling is a blow to the Bush Administration, which argues the inmates have no constitutional rights.**

The shaded parts of the Figure show the key subsentential structures that could be written either to stand alone, or could be moved into combinations with other statements. The parse tree, produced by a probabilistic parser [Cha00] trained on the Penn Treebank, separates the components into clear units. The scope of each unit has to be limited so it is only the *inmates* who have (or do not have) *constitutional rights*, and not the ruling or the

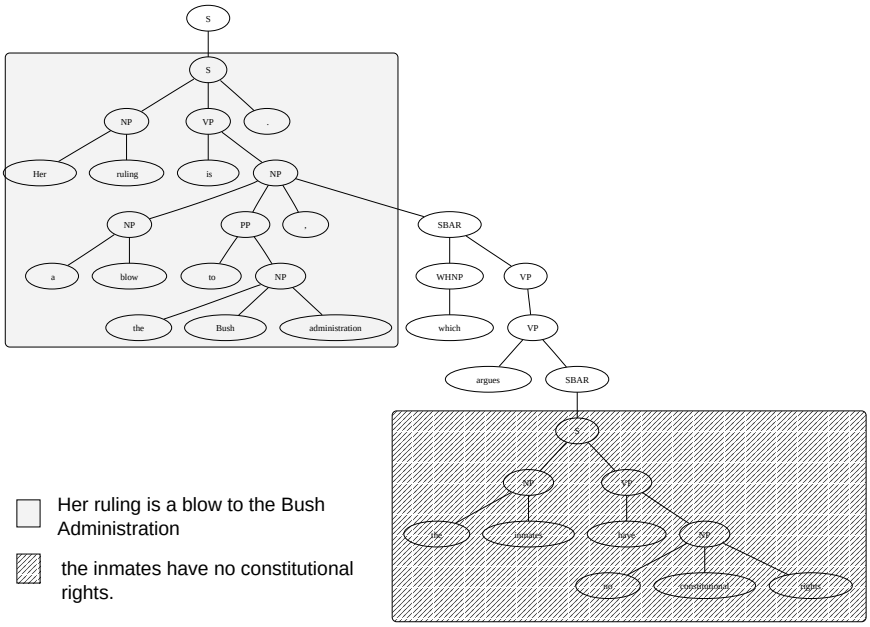

Figure 1.5: A parse tree of a sentence from the BBC article on the U.S. Court decision on terrorist suspects held at Guantanamo Bay in Cuba.

Bush Administration.

But the parse tree provides only a starting point for the analysis. In this sentence, the judge's action is called a *ruling*. Elsewhere in the set of eight articles, it is called an *opinion* and a *decision*. The judge and the court are referred to as *federal* in some places and *U.S.* in others. This adds a semantic and paraphrasing dimension to the new information task. Further, there are two anaphoric references in the sentence above, *Her ruling* and *which argues*, which could easily lead to false classifications. If semantic and reference resolution tools of sufficient power and reliability existed, they could be applied here. But they do not exist. They are active areas of research and either one could be the subject of a thesis. Even the state-of-the-art parser, which achieves about 90% accuracy on test texts from the Treebank, frequently makes errors. At some point, the noise in data will overwhelm even the most sophisticated statistical models or learning paradigms. NLP tasks, like text classification, that can use words as features have no such problems, especially when dealing with news texts, where spelling is highly consistent.

Even when whole documents are at issue, novelty remains a difficult problems. Researchers in information retrieval working on the first-story detection problem (FSD) find it to be harder than similar tasks. A group from Carnegie Mellon [YZCJ02] wrote, "FSD has been recognized as the most difficult task in the research area of Topic Detection and Tracking (TDT), compared to the other tasks like known event tracking and retrospective event detection." In moving from the document level in FSD to the passage level in new information detection, the task grows more difficult. The results at the Novelty Track at the last three Text Retrieval Conferences show that precision scores are rarely much higher than chance – the probability of picking a novel passage at random.

1.4 Contributions

The contributions of this dissertation are:

- Novelty features: I have constructed a set of features useful for novelty detection. They are drawn from two perspectives on the problem, the *Micro View*, which is based on clause structures, and the *Macro View* which looks at whole sentences and beyond sentence boundaries to introduce knowledge of context from previous sentences. Initial experiments and the results at the recent TREC Novelty Tracks convinced me that a bag-of-words-approach based on sentences was going to remain limited no matter what type of modeling was used. The features must be simple enough to be accurately collected but expressive enough to be useful in classification.

- Machine learning: I applied a sampling of automatic learning algorithms to this difficult problem, using an off-the-shelf suite of tools, obtaining the best results from decision-trees and rule-induction. The experiments also covered naïve bayes, support vector machines, nearest neighbors and decision trees. Although training examples were expensive to create and were small in number, the learners were able to cope and arrived at successful hypotheses, tested with 10-fold cross validation.

- Semantic lexicons: I tested a number of ways to use and alter the WordNet hierarchy. I wanted a lexicon that would provide a set of plausible coreference candidates. The

goal was to equate objects across a number of documents, as opposed to strict reference resolution. For example, one article may refer to an aircraft that crashed as a jetliner, but another may call the aircraft an airplane and a third, a flight.

- Information content: I experimented with a new way to measure the information content of words in an effort to find a better way of determining the importance of the terms in a clause. This issue is important in the way many of the features are extracted for determining what passages are novel. Many words are very general, vague or serve some functional purpose and are poor choices for use in the Micro View. These are words like choice, idea or situation. In themselves they carry very little information because they could be referring to almost any kind of entity or event. The $TF * IDF$ metric from the information retrieval community is often used to discount the value of individual terms, but it was apparent that many low-content words are too infrequent to be identified by $TF * IDF$. My effort was based on collecting corpus statistics on the associations between pairs of words, and learning parameters for combining the statistics.

- Multi-document summarization: I implemented DEMS, a multi-document summarizer that uses several innovative techniques, Lead Words, and Cluster analysis. The Lead Words technique is geared to new information, providing a measure of importance for words that have not yet been encountered. It is based on comparing distributions of words that tend to appear in the leads of news stories and those that appear anywhere in news stories. Cluster analysis determines the topology of the cluster. If a cluster is bound to a central event or entity, the summary should focus on that, and ignore peripheral articles. But if the cluster is more diverse, perhaps a discussion of a kind of event, the summary should try to say a little about each article.

1.5 Outline of the Thesis

Related Work Chapter 2 surveys research by others in the area of new information detection, which was mostly in the TREC Novelty Track, and in ancilliary areas like the kind of semantics I applied to this problem.

Corpus Chapter 3 describes how I compiled and annotated a development corpus of 31 pairs of news articles. Each pair was on a particular event and was annotated by two people not involved in this research. After their initial markup, they negotiated the differences in their judgments to produce a set of 2,400 examples.

Syntax Chapter 4 discusses the rationale for employing a full-scale parser; in this work, I show why a sentence-by-sentence comparison does not work well in the new information task, and why the *Micro View* is appropriate. . To recognize clause structure with a high degree of accuracy, I use a probabilistic parser.

Semantics Chapter 5 covers the extension of the WordNet lexical database that I use to create equivalence classes for common words, for example, in order to equate the mention of a *car* in one article with a *vehicle* in another. In addition, the chapter shows how a named-entity recognizer is used to find a single canonical name for people, organizations, locations and other named objects, and I describe efforts to measure the content a word carries.

Context Chapter 6 is about my exploration of the *Macro View* at the last Novelty Track evaluation and why it is advantageous to look at the previous passages in determining the novelty of the passage currently being examined.

Machine Learning Chapter 7 reviews the types of machine learning algorithms I tried and my choice of rule induction with Ripper [Coh95]. The experimentation was done with the Weka suite of learning tools [WF00].

Evaluation Chapter 8 deals with the different features used in learning, and the final selection of model, and it details the results obtained.

Summarization Chapter 9 describes the DEMS summarizer and shows an example of how **NIA** works with DEMS used to present the output.

Conclusion Chapter 10 gives my conclusions about the work, discusses the limitations of what I've done, and directions for future work.

Chapter 2

Related Work

This chapter will present an overview of research in new information, almost of all of which has been done in conjunction with the TREC Novelty Track, and recent work in multi-document summarization and relevant work in semantics.

2.1 TREC

The task in the Novelty Track is related to first story detection, which is defined on whole documents rather than on passages within documents. In Task 2 of the Novelty Track, the inputs are the set of relevant sentences, so that the program does not even see the entire documents.

At the recent TREC, Dublin City University achieved some of the top scores by comparing the words in a sentence against the accumulated words in all previous sentences [BBC+04]. Their runs varied the way in which the words were weighted with frequency and inverse document frequency. Like my approach, which I call SumSeg, theirs follows from the intuition that words that are new to a discussion are evidence of novelty. Their weighting of words in part by inverse document frequencies is also similar to my use of document frequencies to help identify low-content words and prevents them from triggering a false novelty classification. But my approach distinguishes between several kinds of words, including common nouns, named persons, named organization, etc. My approach also incorporates a mechanism for looking at the context of the sentence. I also learn the various weights and

thresholds used by the system to target either precision or recall.

Both the Dublin system and mine are preceded by the University of Iowa's approach at TREC 2003. It based novelty decisions on a straightforward count of new named entities and noun phrases in a sentence [ESL+03]. I was struck by their performance, as the Iowa system was the only one to push precision beyond a narrow range that year. In 2004, the Iowa system [EZB+04] tried several embellishments, one using synonyms in addition to the words for novelty comparisons, and one using word-sense disambiguation. These two runs were above average in F-measure and about average in precision. Their remaining three runs were based on a vector-space model with an interesting twist in that the novelty threshold was computed dynamically. Precision was lower on these, and recall varied.

The University of Massachusetts system [AJAC+04] closely resembled my combination submission, mixing a vector-space model with cosine similarity and a count of previously unseen named entities. Their submission used a similarity threshold that was tuned experimentally, while mine was learned automatically. They identified unseen named entities, while I identified unseen nouns and verbs in addition to named entities. Their combination operator was also different. The UMass group used the union of the two subsystems while I used the intersection. In earlier work with the TREC 2002 data, UMass [AWB03] compared a number of sentence-based models ranging in complexity from a count of new words and cosine distance, to a variety of sophisticated models based on KL divergence with different smoothing strategies and a "core mixture model" that considered the distribution of the words in the sentence with the distributions in a topic model and a general English model.

A number of groups have experimented with matrix-based methods. In 2003, a group from the University of Maryland and the Center for Computing Sciences [CDO03] used three techniques that used QR factorization and singular value decomposition to analyze word-by-sentence matrices. The University of Maryland, Baltimore County, worked with clustering algorithms and singular value decomposition in sentence-sentence similarity matrices [KSC+03]. In 2004, Conroy [Con04] tested Maximal Marginal Relevance [GMCK00] as well as QR factorization.

The information retrieval group at Tsinghua University used a pooling technique, grouping similar sentences into clusters in order to capture sentences that partially match two

or more other sentences[RZZM04]. They said they had found difficulties with sentence-by-sentence comparisons.

The Institute of Computing Technology, the Chinese Academy of Sciences [ZXB+04] used word overlap, as well as several similarity computations similar to Maximal Marginal Relevance, and information gain. They dropped a technique used in 2003 of varying the number of novel sentences they accepted depending on the ordering of the input documents [SpZ+03]. That would make it less likely for sentences to be drawn from the later documents.

Meiji University embellished pairwise similarity calculations with co-occurrence data from a background corpus. It restricted the novelty comparisons to a time window for the publication dates and included an inverse-document-frequency term in scoring sentences [Uni03].

An interesting approach at TREC 2002 was done by a group at CMU [CTOZC02], which used a graph-matching algorithm to compute similar structure between sentences. They used WordNet to identify synonyms but this was limited to the synsets. In contrast, I used hypernyms (more general related terms) and hyponyms (more specific related terms) and experimented with several ways to augment the WordNet dictionary.

2.2 Summarization

My approach is also related to multidocument summarization, but most summarization systems are based on locating similarities between documents [MKH+99], [HMM+01] and [Mar01]. A group at CMU [GMCK00] uses cosine similarity of vectors in the maximal marginal relevance(MMR) algorithm. They seek to eliminate redundancy from their summaries with a measure similar to Allan's novelty detector. My handling of similarity differs from these because it seeks to equate different words that refer to the same underlying concept and by its use of clause boundaries rather than sentences to establish a link between terms. Yet similarity is only the beginning of the computation. I do not compare similarity of clauses, but the document-wide coverage of the concepts contained in the clauses. Thus a clause composed of strongly matched concepts does not receive a high difference score, even if there is no similarly structured clause in an earlier document.

In other work, a graph representation of several relationships between words is used by researchers at the Mitre Corporation to find similarities and differences between pairs of articles [MB97, MB99]. They recognize that sentences cannot be examined independently, without reference to other sentences in the same article. The thrust of their summarization is user-oriented in that the summarizer is responding to a user's query. Salience is semantic closeness to the query, but the authors note that a $tf * idf$ weighting scheme could be substituted. They recognizes differences among the input articles by the set of unique words in the segments containing the salient passages.

Dragomir Radev [DRRB00, RJB00] seeks to define a subsumption relation between sentences in building multi-document summaries. He computes a redundancy penalty, which is based on word overlap between a pair of sentences.

A group from Cornell and Cogentex is looking at the related problem of "discrepancy detection," in particular those of numerical differences [WCN+01]. Theirs is a much finer-grained task, but limited in scope by concentrating on numbers, a much restricted problem.

The summarization component I will use, DEMS is closest in spirit to NEATS [LH01], a summarizer that was also tested at DUC. NEATS uses *topic signatures* to try to obtain a better measure of frequency. Topic signatures are based on co-occurrences in a background corpus to give more weight to terms that are often seen in the same context. DEMS tries to group semantically equivalent words into sets, called Concept Sets, in determining the salient entities.

Several recent developments in multi-document summarization parallel a key idea in **NIA** – that the appropriate unit of information is not a sentence. In Multigen [MKH+99], developed at Columbia, redundant subsentential phrases are exchanged in *information fusion*. The trend in summary evaluation is also moving away from direct sentence comparisons. The ROUGE tool [Lin04] for automated evaluation at the Document Understanding Conference is optionally using "skip bigrams", which are pairs of words appearing in the same "Basic Element", obtained from parse trees of sentences. In addition, Van Halteren and Teufel [HT03] propose using "factoids" as an atomic unit of information, and Nenkova and Passonneau [NP04] offer "Summary Content Units". All three are used for evaluation of summaries, not the construction of summaries, and the last two are done by humans.

2.3 Semantics

My hypothesis requires some semantic information to make the novelty computations more powerful, but this need is closer to a surface analysis than is found in much of semantics research. My primary purpose is to obtain reasonable coverage of equivalent words – as a thesaurus would. I start with WordNet and then try to add to it and combine it with other resources.

2.3.1 WordNet

WordNet has become a widely used lexical resource in the Natural Language Processing community and many have proposed extensions, modifications and uses for it. Most of the proposals either want to add information to the database [AAMH01] or to perform complex tasks like word-sense disambiguation with the the existing entries [HMM99, GPDR01, IH01]. The effort that seems closest to what I am proposing is the transformation of WordNet into a coarser-grained dictionary to reduce polysemy proposed by Mihalcea and Moldovan [MM01]. The authors target the tasks of word-sense disambiguation and machine translation. While such a modification would be useful in my research, the problem I have found goes beyond that. First of all, I find that I must look beyond the synsets to the hypernyms and hyponyms in order to locate references. If I use only synsets, the system misses too many; if I use all hypernyms and hyponyms, it frequently goes off track. Many of the difficulties of using WordNet as is are discussed by Resnik [Res99], who proposes a measure of similarity based on the information content of nodes in the WordNet hierarchy. His aim is to establish a metric to guide such tasks as resolving syntactic ambiguity. This is different from the requirement that I have in establishing links between words sometimes many paragraphs away from one another and sometimes in different documents.

The need for some combination of existing lexical and semantic resources was described by Zickus and Mccoy [ZMDP95] in their proposal to build a lexical database from a variety of existing lexical resources, including WordNet.

Word similarity has also been extensively studied by psychologists, like Tversky [Tve77] who suggested that the similarity between concepts could be measured by the features they

have in common and those on which they differ. His work leads to many of the unsupervised methods for automatically obtaining semantic lexicons like the one discussed below.

2.3.2 Automatic Methods for Obtaining Semantic Data

The simpler versions of unsupervised methods analyze co-occurrences of words within some window, ranging in size of 2 or 3, to entire documents. The underlying idea is that similar words will appear in similar contexts. This technique is similar to relevance feedback in Information Retrieval, and has been applied to multidocument summarization where Lin and Hovy used *topic signatures* [LH00] to choose sentences for inclusion. Others have used singular value decomposition in a similar way [JJDM01, SJM+02, DDJ+03, JJJD04]. Intuitively, these methods are problematic, first because they are inherently noisy, and second because they are as apt to find collocations ("computer security") rather than equivalence classes ("computer → machine"). This kind of analysis often captures relatedness rather than referential potential and is a much looser notion than what is required for new information detection.

More elaborate systems often use some kind of syntactic information to base probabilistic calculations of similarity. In one of the key early papers on the subject, Pereira [PTL93] clustered verb-object pairs to determine similarity, using information theoretic measures of similarity. Hatzivassiloglou and McKeown [HM93] group adjectives by their meaning, using syntactic patterns, linguistics tests and nonparametric statistics. I tried to use the inherent coherence in document to collect co-occurrence statistics from sequences of subjects [SM00] in an effort to discover information about likely coreferring expressions.

Later, Lin [Lin98a] used a large number of syntactic relations that he found automatically with his rule-based parser MINIPAR [Lin98b], and determined similarity by computing the mutual information of pairs of words in various syntactic relationships, like subject-verb, verb-object, with each other. The statistics were drawn from a 64-million-word corpus of news, from the Wall Street Journal, the San Jose Mercury and the AP. This dictionary is publicly available, and as such, provided a way to experiment with the results of unsupervised methods and maintain the focus of this thesis on the novelty issue. Continuing my own efforts at thesaurus-building would easily have filled another dissertation.

2.4 Word Importance

In trying to determine importance, my technique borrows from some recent work in the machine learning community, where training material is found that is already partially annotated or structured in some way to provide an adequate number of positive training cases, without the need for expensive markup of data. In an information extraction experiment Mark Craven [Cra99] used what he called "weakly" labeled data to reduce the cost of annotating a training corpus. He was seeking a way to map medical texts into a structured data base. He used a database that contained links to related text articles. Thus he could automatically collect relevant articles, reducing the effort to prepare his training corpus.

Another group working on information extraction at CMU, Seymour and others, sought to build a database of information about computer scientists from "distantly labeled data" composed of the header information on research papers. They reported high accuracy with Hidden Markov Models trained over this prefabricated data [SMR99]. Ellen Riloff learned textual-syntactic patterns for information extraction by comparing two corpora, a target containing the information she was interested in, and the other a general corpus [Ril96]. The idea is that patterns of specialized words and syntactic structures will show up in greater numbers in the target corpus than in the general corpus.

My experiment is similar to all three in that I am considering the first paragraphs of news articles as a preselected corpus of *important* and *interesting* information. But my system differs from the two CMU groups, which use a more structured kind of data, and from the Riloff work, which is seeking to find very specific patterns from a very loose corpus.

An alternate approach for using unlabeled data comes from Rie Kubota Ando [And04], who used singular value decomposition on matrices built from vectors of syntactic feature values, such as subject-verb and verb-object relations to categorize common nouns into several broad categories such as as persons, locations, facilities. Several hundred seed nouns from the target classes are given to the system, which then searches for the best features for distinguishing the different targets. When the dimensionality of the matrices are reduced, the best features are retained. The goal here is distinct from mine as there is no intuitive link between word importance and syntactic patterns.

Chapter 3

Corpus

The experiments and development of the system for this thesis required a suitable corpus of annotated texts that I had to construct in order to begin the work. The texts needed to be clustered by topic and similar enough to provide interesting comparisons with sufficient differences in details. I needed articles about the same specific event that they were published at about the same time.

I decided to explore the problem by examining pairs of related articles of which one article would be considered the background and the other a new report, possibly containing new or novel information. By comparing the pairs, I would not lose generality, since the old article simply represents all the *background*, the entire collection of previously seen statements on the event. The problem would not change if the *background* was drawn from 1 article, or 10, or 1,000. At the same time, the new article represents the entire collection of unseen statements that must be read against the collection of previously seen statements from all the old articles. It could be a single document, or a cluster or documents ordered in some way, or an online stream of documents, such as those from a newsfeed.

NIA is intended to operate in an on-line situation, dealing with large amounts of text as in a stream of news reports. In NewsBlaster, **NIA** functions in conjunction with a tracking service [MBC+03], which follows events over a period of days. With each day's addition of current news, offers users a graphical view of developing topics. The graphs are redrawn and **NIA** produces a new update. The tracking functions are responsible for dealing with the burstiness of news events observed over a long period of time [Kle02].

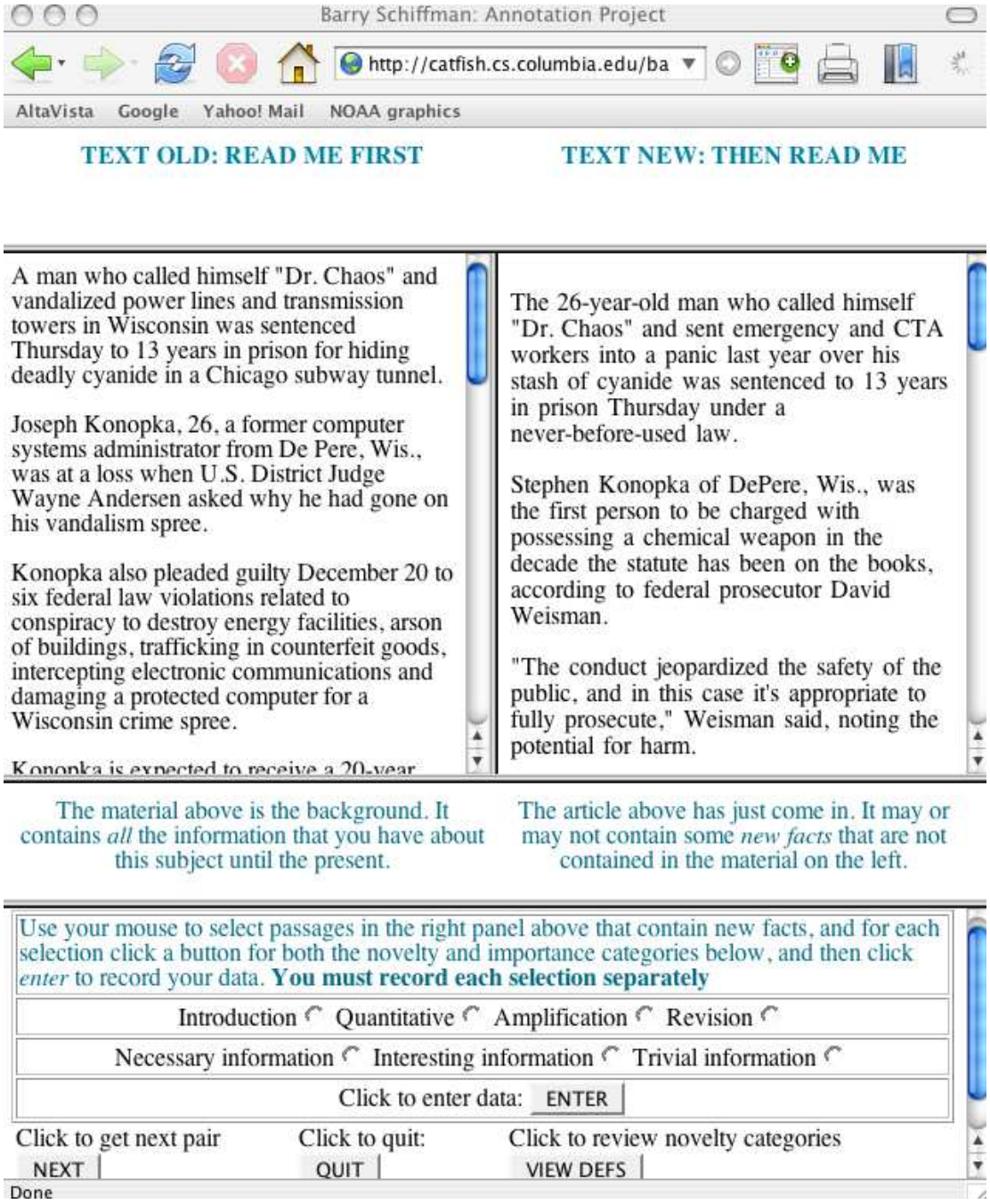

Figure 3.1: Example of the annotation interface, displaying a pair of articles and the form for the annotators. The article on the left is the *background* and the article on the right is being scanned to see if it contains any new information.

By using pairs of documents, I also intended to make the most efficient use of my annotators' time and effort. Reviewing documents myself showed that the annotation would be time consuming under the best circumstances, and the demand on the annotators' time would increase at a rapid rate. For each statement read in the Current Document, all of the background would have to be reviewed – a $O(n^2)$ task in terms of complexity. Of course, I needed to have sufficient inter-annotator agreement in order to judge what strategies were working and what weren't.

The goal in building the development corpus was to create manageable, realistic trial inputs for a summarization system that would output updates or bulletins on the news.

The challenge was to obtain the corpus just described using limited resources.

The corpus had to meet two main requirements:

1. The annotations had to be flexible. I wanted to be careful that the annotation setup did not restrict how the actual classifier could be constructed. Giving the annotators parsed text and asking them to identify components might have locked me into to using the units which that parser recognized.

2. The article pairs had to be on topic, with some overlap so that the comparison would be meaningful in terms of the new information task. A pair of unrelated articles would be trivial since all of the unseen article would be novel, and a pair of near duplicates would likewise be trivial. There is a continuum between these two extremes, and the proportion of new information varied from 20% to 80%.

The first requirement was dealt with by having the annotators choose unlabeled spans of text in a simple web browser interface. The selected words were then mapped to a parse, and in fact, I was able to change from a fast, finite-state parser to a full-scale probabilistic parser by changing the script that performed the mapping. The mapping is done by computing the proportion of selected words that appear in each clause, and testing to see if it equals or exceeds a threshold – which was set at 50%. The news genre offered a plentitude of material, a choice of different articles of ordinary, well-formed English text in a wide range of topics in order to meet the second requirement. The details of the annotation will be covered in Section 3.3.

3.1 Comparison With TREC

My use of pairs of articles is in contrast to the task in the TREC Novelty Track, where the inputs to the novelty-detection system are lists of sentences that were taken out of context. These sentences are the ones the NIST assessors deemed relevant to the given topic. The proportion of relevant material to the input documents in the Novelty Track varied greatly, from 2.85% to 70.53% in 2004. Looking at the novelty filtering task alone, the proportion of new material to all the relevant material ranged from a low of 12.82% to 91.49%. The document clustering in the Novelty Track was done automatically by an Information Retrieval system, and reflects the reliability of the system and the distribution of the underlying corpus, which was about three years of newswire by three services, the New York Times, the Associated Press and Xinhua [1], as well as the choice of topics. Often the articles relevant to a topic span a period of two or three years. For topic 55 (See Chapter 1), which concerns nuclear weapons tests by India and Pakistan during a three-week period in May 1998, the articles were published from May 1998 through May 2000. Thus many of them treat the event as distant in time. In NewsBlaster, much of this difficulty is removed because the news organizations that publish the web sites NewsBlaster crawls are restricting their material to the events that are current on any particular day.

By removing the contexts, which can change greatly during the time the articles cover, the Novelty Track task imposes a limitation on the kinds of analysis that could be done, unless the system also goes back to the original inputs. This, however, creates a dilemma because the assessors were to read the list of relevant sentences and not the original articles, but they obviously may or may not remember some details and may or may not look other details up. Although my hypothesis states that contextual features are important, a strategy using contextual features will be limited here to sentences that are contiguous in the original articles. In effect, the TREC task encourages direct sentence by sentence comparisons.

For example, below is a sequence of three sentences from one of the relevant articles in the TREC Topic 55. The Sentence 7 and Sentence 9 were selected as relevant, but Sentence 8, which lies in between, and is shown against the gray background, was not:

[1] The AQUAINT corpus.

Document XIE19980529.0283, Sentence 7: Pakistan carried out five nuclear tests on Thursday in the wake of the same number of tests by neighboring India earlier th is month, Pakistani Prime Minister Nawaz Sharif said in a televised address to the nation.

Document XIE19980529.0283, Sentence 8: The Italian news agency ANSA said Italian Foreign Minister Lamberto Dini, in Luxembourg attending a foreign ministers' meeting of the North Atlantic Treaty Organization, said "strong concerns" were aroused by the fact that another developing country conducted nuclear tests after India.

Document XIE19980529.0283, Sentence 9: He urged Pakistan and India to halt their nuclear tests which, he said, constituted a "serious reality" for the international community, and called for an early and thorough implementation of the Nuclear Non-Proliferation Treaty.

Clearly, Sentences 8 and 9 go together. The article itself contains the first reference to the Italian government's reaction, and the first reference to Foreign Minister Dini. But without Sentence 8, it is impossible to guess who urged the two countries to halt the testing. Possibly, a human reader would guess that it is not the Pakistani Prime Minister who is urging a halt to the tests.

Although NIST prepared a far larger amount of data for the Novelty Track, the evaluation used the judgments of a single human assessor in the 2003 and 2004 evaluations. A second human assessor evaluated the data for task 1, which combined retrieval and novelty task, but NIST did not calculate statistical measures of agreement[2]. Instead, the second assessors were used as a upper bound for performance in Task 1. Interestingly, the second assessors were far better than all systems in precision, both in the relevance phase and in the novelty filtering phase, but substantially below the automated systems in recall. The second assessors could not be used in Task 2 because they scanned their own relevant selections.

[2]The primary judgments were always those of the authors of the topics; the topics were divided among all the assessors.

3.2 Pilot Markup

In light of the difficulties with the TREC data, I collected a group of article pairs early in 2002 – soon after embarking on this work – and conducted a pilot annotation. The pilot revealed numerous problems. I asked 21 people who played no part in this research to look at four pairs of documents. The people had varied backgrounds. Some were graduate students involved in computational linguistics research; others had varied amounts of education. Most were native speakers; some were not. Ages varied from student-aged to middle-aged. All were asked to mark passages that contained new and important information. They performed the markup task via a preliminary web interface that recorded their selections and allowed them to write in comments. After looking at the annotations, and reading the annotators comments, it was clear that decisions about novelty and importance conflicted. Each determination is to some degree subjective, making agreement more difficult to achieve.

The annotators were all volunteers, and agreement among them was poor. Eight of the sets were given to two annotators to test agreement and the average Kappa coefficient was 0.24, well below levels generally considered to be acceptable – 0.6 or more. Examination of the annotations showed that different people had different notions of what constituted difference between statements, with decisions often turning on what might be legitimately inferred by the texts. This was a particular problem because the corpus was drawn from the news of that time, which was shortly after the Sept. 11, 2001, terrorist attacks, and the news writers often assumed that general readers were entirely familiar with the events. For one example:

> "Tough call, Osama bin Laden was mentioned in the first article, and I think the author assumed the reader already knew he was the prime suspect, but the first article didn't say it, and I'd call it important."

And, another similar comment:

> "The idea is deducible from first article but in the new article is directly expressed."

The annotators were given the opportunity to comment on their selections, but since they were volunteers, I did not require them to do so.

I also chose articles that were in some cases too far apart. For example, one pair was on the mayoral election in New York City, one a day before the vote, and one a day after. The central difference was clear – the voting was completed and a winner was declared. But the article pair showed that the changed circumstance in the world was accompanied by a shift in the reporters' focus so that indeed large portions of the documents were orthogonal. This raises an important question: On the one hand, a system that provides updates to users would ultimately have to cope with such messy inputs, but on the other hand, I needed to break down the task into subproblems that could be tackled one by one. First, it was clear that I ought to ask the questions of importance separately from those of novelty. It also seemed clear that I had to choose the pairs of articles more carefully. I decided to make the assumption that the system would have a good clustering system providing clean input, so that the pairs would be indeed closely related, and development could proceed on examining the inputs in detail.

One other major issue surfaced in the markup. Occasionally, the annotators seemed to be inattentive and careless. The amount of time it took to do the four pairs varied considerably from about 20 minutes to two hours. The faster ones tended to pick one or two statements as novel and to leave it at that.

3.3 Experimental Setup

The problems of the pilot study led to a number of fundamental changes in how I built the development corpus. These changes were primarily aimed at obtaining a more consistent, higher quality set of examples of novelty. I decided to recruit journalism students at Columbia University to do the annotation and to pay them a competitive rate for part-time student work ($10 an hour). The journalism students would be good readers, and they would be thoughtful about what makes an utterance *new* compared with previous statements, since novelty is an important element of what goes into the quality that journalists call "news judgment."

And I decided to require them to make two passes over the pairs and then to have them negotiate any remaining differences of opinion. In the first pass, they used an *HTML* interface available over the *World Wide Web* (See Figure 3.1). They used the mouse to select spans of text that contained new information. After both of the annotators assigned to a pair of documents finished, I automatically rewrote the web page with a color code, with blue showing the areas that both of them agreed were novel, and with red showing the areas that only one of them marked as novel. Thus, areas left with the white background were areas that both agreed were not novel. In the second pass, the annotators were asked to review the selections, focusing on the areas where there was disagreement and to reconsider them. Several of the annotators said that they had expected the second pass to be fast and easy, but that they found it as difficult and time-consuming as the first pass. I surmise that there might have been a certain amount of carelessness in the first pass, and that the coloring of the second pass focused the annotators' attention on specific passages. In any case, a large amount of the initial disagreement disappeared in this second pass, simply by making the annotators review their own work. Then, they were required to resolve the few remaining differences.

On the first five pairs of articles, our annotators disagreed on 48 fragments – contiguous runs of words marked novel – almost half of the fragments initially marked. But in this second, reconsideration, pass, 35 of those disagreements were resolved – simply by having the annotators take a second look. At this point, only 13 disagreements were left. To resolve these remaining points, they began an exchange of e-mails; in this case 9 points were resolved in the first exchange, and the last 4 in the final exchange.

Figure 3.2 shows passages from the pair of articles on the actor Robert Blake, who was accused of killing his wife. The passage on the left is from the *background* and the one on the right is from the new article. One annotator marked the bold face section on the right, and the other didn't on the first reading. The two remained in disagreement after the reconsideration. The instructions told the annotators to judge novelty separate from importance, and asked them to categorize both the novelty and importance of each passage.

But the annotator who favored novelty, argued in the negotiation phase, "The 1st article didn't say this was the first time he had appeared in court since posting bail." With that,

Before actor Robert Blake pleaded not guilty Thursday to killing his wife, **he spoke to reporters for the first time since he was released from jail two weeks ago**, thanking the judge who set his bail for "saving my life" and calling himself the "luckiest person you will ever meet."	**Appearing in court for the first time since he was allowed to post bail March 14**, a healthier looking Robert Blake pleaded innocent to murdering his wife and waived his right to a speedy trial until October, when proceedings are likely to begin.

Figure 3.2: Passages in the box on the left were drawn from the **background** article on the actor Robert Blake. Passage in the right box was marked as novel by one annotator and not the other.

his counterpart agreed. There is no reason to suppose that a defendant only speaks to reporters when he appears in court[3].

My negotiation strategy is very close to what was done several years ago at Columbia in developing Simfinder [HKH+01], a tool for determining similarity between paragraphs. A data set of matched pairs of paragraphs drawn from the Reuters part of the 1997 TDT pilot corpus was created in a two-stage process. First, each pair of paragraphs was judged by two human reviewers, working separately. The reviewers were asked to determine whether each pair of paragraphs contained "common information" – whether they refer to the same object either performing the same action in both paragraphs or described in the same way in both paragraphs. The reviewers were then instructed to resolve each instance about which they had disagreed. The annotators were able to resolve their differences and come with a single label of similar or not similar when they conferred after producing their individual judgments.

The negotiation phase in my annotation is also similar to the adjudication used in several important markup projects, notably the creation of the Penn Treebank [MSM93] and more recently the PropBank [KPM02], a project at the University of Pennsylvania to enrich the Treebank with information about verb subcategorization. The researchers used

[3]As far as importance, the annotator who marked it novel, also considered it necessary, but the importance categorizations were not dealt with in this research.

an adjudication procedure to improve accuracy. Most of the annotators are undergraduate linguistics students, who are given several days of training, but who come to the task with some expertise. The investigators note that interannotator agreement is relatively high, but they used a small team of highly trained linguists to perform the adjudication "to catch and correct discrepancies."

To keep the task simple, I displayed the pairs of articles side by side. I didn't try to carefully define *novelty*, but left it to a commonsense notion. However, I sought to take advantage of some expertise in making these kinds of judgments by recruiting among journalism students. The notion of picking out fresh information is at the heart of what working journalists call "news judgment." I did not define any particular structure to be marked – like sentences or clauses. Each word was indexed so the marked segments could be retrieved later.

The general instructions on the opening page for the annotators described the task like this:

> You will be shown pairs of news articles side by side. The article on the left will represent all the background you have on a particular event – for example, a criminal case. The article on the right will be a new article, which may be loaded with new information or which may essentially repeat the information you already have. You will highlight, using your mouse, segments of the article on the right that contain new information, without considering how important.

By separating the two judgments I avoided clouding the question of novelty from the more subjective notion of importance. For each selection, the annotators were forced to categorize their selection in two ways: *novelty* and *importance*. I gave them these choices to respond to the two questions.

Type of Novelty Introduction, Quantitative, Amplification, Revision

Degree of Importance Necessary information, Interesting information, Trivial information

Each pair of articles was on a particular event. I aimed to find articles that largely covered the same ground but differed in detail. The articles were usually published in short

spans of time, often within a day of one another, and often published by different news outlets. I favored articles about more obscure events to force the annotators to rely on the texts rather than to draw on their own knowledge of the events of the day. Some examples are:

- the announcement of Marlon Brando's divorce settlement
- the marriage of a convicted murderer on death row
- the lawsuits filed by America On Line against spammers
- the proceedings against the television actor Robert Blake.

For the first few pairs, I chose some topics by hand and searched for available articles in the Lexis/Nexis database. Then, I selected NewsBlaster sets and automatically reclustered the articles, choosing the pair that was closest in content without being duplicates. In reclustering, the system builds an abbreviated similarity matrix for all the documents. By abbreviated I mean that similarity is computed only on the n most frequent words in the document, and the pair of most similar documents were taken for the annotation. The instructions to the annotators can be seen in Appendix B and the texts of the articles in Appendix C

The reason for reviewing the documents by hand was economy. An alternative would have been to use an automated choice of documents, but document similarity is a much different problem, and automatic clustering algorithms make a fair number of errors. While relying on automatic document clustering would have provided a real-world dimension to the my task here, it would have been expensive to obtain annotations for documents that are obviously orthogonal to each other.

To identify the old and new documents in the later chapters, the background will refer to the old document or documents, i.e. those that have already been seen. The processing of documents is cumulative. As each new document, the current document, is scanned by the program, it is examined for novelty and then added to the background. I assume that information in a particular document was not repeated within that document, so that the current document is only added to the Background.

In this exercise, one important assumption is made: that *novelty* is largely an objective judgment, on condition that it is considered without respect to importance and that it is based solely on the facts that appear in the two articles (or in the real-world conditions, as in NewsBlaster, the two sets of articles. The instructions to the annotators, and the process of negotiation, encouraged this kind of literal-mindedness. The annotators were told to try to divorce themselves from any knowledge they had, and to try to base their classifications only on what they saw in the two articles. In light of the difficulty of the task, it is a good starting point.

The result of the effort was a reliable set of marked 2,023 clauses from the new documents in each pair, and of this total, about 61% were classified as novel. Although the effort was labor intensive and expensive, the pairs were particularly helpful in the development stages of system building. By operating on pairs, it is feasible to review program's decisions, particularly in dealing with the syntactic and semantic issues discussed in Chapters 4 and 5. In Chapters 7 and 8, the training set proved to be sufficiently large for machine learning experiments.

Chapter 4

Syntax

The definition of new information detection presented in Chapter 1, like that of the TREC Novelty Track, implies a need for some degree of language understanding. My hypothesis claims that a substantial amount of linguistic information will be necessary to answer the question of whether a statement is novel – that is, whether or not it contains any information already seen in some previous set of statements. I view the work in this thesis as an investigation into how much understanding is sufficient for reasonable results. For example, contrast my task with that of establishing similarity. Suppose two sentences have a 90% simple word overlap, a purely surface comparison. They are almost certainly *similar* but they could easily convey opposing opinions, contradictions, or new developments. This is illustrated by some fragments from a Novelty Track set[1]:

The first pair of sentences about British Airways and Air France are short, clipped simple ones. They were taken from an Associated Press article on Aug. 3, 2000, about a week after the tragedy. In both sentences, how easy it would be to change the meaning with a single word.

> British Airways **resumed** its Concorde flights a day after the crash. Air France Concorde flights were immediately **suspended**.

Two weeks later, on Aug. 16, the following sentence was reported by the Xinhua news service. They are more complicated and contain more background.

[1]Set 69 in the 2004 evaluation, on the crash of the Air France Concorde in 2000

> Air France has **suspended** the flight of its remaining five Concorde planes after one of such plane crashed on July 25 near Paris. British Airways **suspended** the flight of its Concorde planes on Tuesday, nearly three weeks after the July 25 crash in which 113 people were killed.

The goal in this thesis is to be certain to identify the change in the British Airways situation during these two weeks – the actual shift in the verb from *resumed* to *suspended*, while recognizing that there is no new information about Air France.

The Xinhua sentences contain additional details which the Associated Press articles made clear by the context. These details make a direct sentence-by-sentence comparison difficult, especially when the sentences are represented as bags of words without any consideration to structure.

At the center of my approach, I examine smaller units of text than the sentence, namely clauses containing a verb and its arguments. Clause boundaries are found by a probabilistic parser, but I tried using a minimal amount of analysis of the argument structure. In linguistics, these structures are often labeled as S and $SBAR$ nodes, indicating a sentence-like structure, and an English sentence may be composed one or more of these, often with some clauses embedded in others. Writers embed background information in other clauses. Some important changes about the event or the topic can be surrounded by older, already known details. The two Xinhua sentences above both contain a retelling of the basic event – the plane crash.

I view clauses as atomic units that each carry a single statement. Writers are free to arrange and rearrange clauses without altering their meaning. What I sought to do was to reduce a document to a list of atomic statements that can be easily compared to a set of previously seen statements. To continue with the Concorde example, we may have many clauses in the previously seen documents that contain references to the accident, so their presence alone would not indicate novelty.

Here is a passage from a New York Times article about the accident:

> British Airways PLC **grounded** its fleet of Concorde supersonic airliners Tuesday after investigators said they would urge the withdrawal of the airplane's

> airworthiness certification following the fatal Concorde crash near Paris last month.

This is a fairly complicated sentence, but at 33 words, it is a typical length, though a bit higher than average. It contains a reference to British Airways suspension, and this reference must be isolated from the rest of the material in order to determine novelty.

Figure 4.1 shows a full parse of the passage. The material about the status of the Concorde is enclosed and highlighted. It's easy to see that the statement about the aircraft's status is only a fraction of the sentence, one element in three. The sentence also includes information about *airworthiness* and a condensed mention of the where and when and nature of the crash.

I am making the assumption that the content words within any clause are related to one another in some way. Under my hypothesis, I will be able to decide novelty by using these unordered word pairs, which I call *co-occurrences*, along with other features. First, there are a number of important details to take care of.

With an appropriate decomposition of the texts, the passages can be represented as a collection of relationships. Table 4.1 shows how the clause in Figure 4.1 can be broken down.

	Brit Air	Concordes	ground
Brit Air	-	R	R
Concordes	-	-	R
suspend	-	-	-

Table 4.1: An R in a $cell_{i,j}$ indicates that $term_i$ in row_i is somehow in some meaningful relationship with $term_j$ in $column_j$. These are the co-occurrences that the system extracts.

The table shows, for example, that there is some kind of relationship or connection between *British Airways* and the verb *suspend*. The table shows only that a relationship exists, and does not specify the relationship R. The co-occurrences mean only that the two terms co-occur in some clause in the text. No effort is made to pin down the grammatical relationship. It could be that British Air was suspended from some activity, or that British

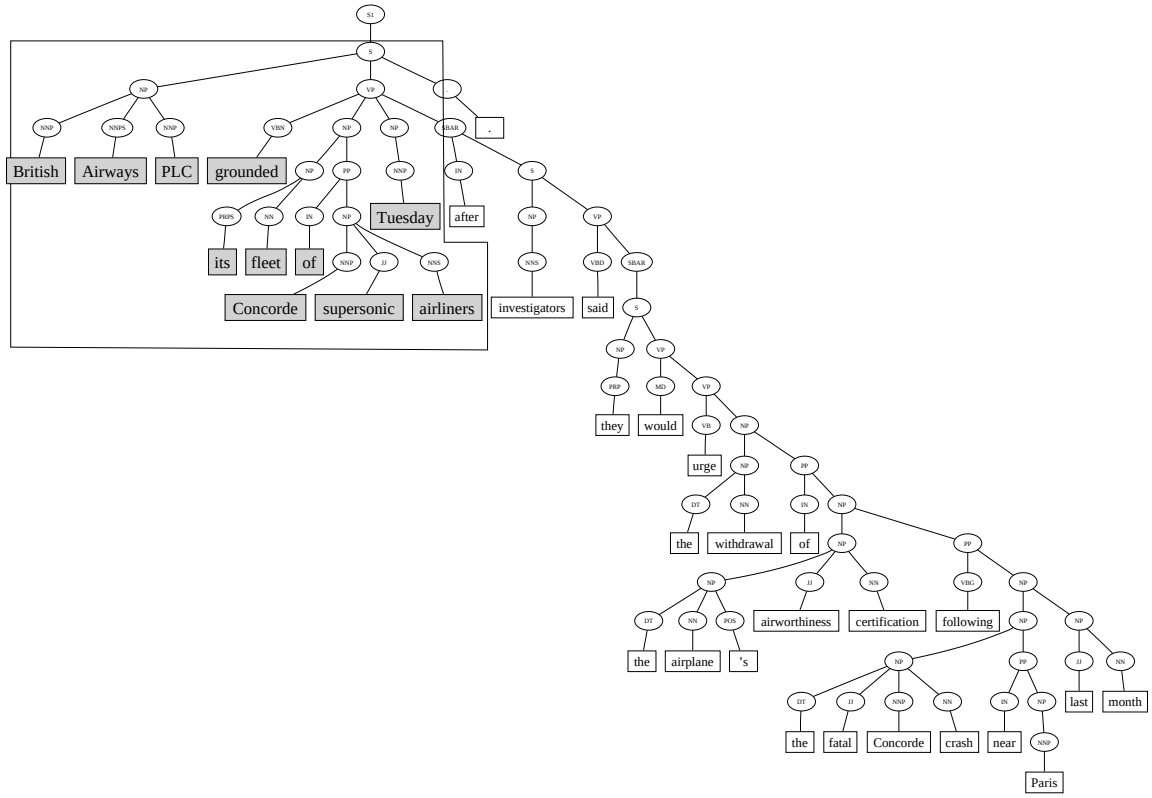

Figure 4.1: A parse tree of the New York Times passage in the style of the Penn Treebank notation. The enclosed portion shows a clause that could be a single sentence, or affixed to another sentence or embedded into another sentence without changing the meaning – or even the emphasis. The parse was obtained automatically by Charniak's probabilistic parser.

Air suspended something. Thus, the work in this thesis goes beyond a bag of words in that a detailed syntactic parse is taken of the input text, and co-occurrences, or pairs of words, are recorded, based on the parse, but this is far from anything like natural language understanding.

For the sake of simplicity, this table reduces the noun phrases to a single concept so that *a fleet of Concorde supersonic airliners* is represented as a single entity, *Concordes*. This is a program option that will be discussed below. In any case the number of potential relationships is quite large. For any n words in a clause, there are:

$$O(\binom{n}{2})$$

but in reality n is expected to be small almost all of the time.

The co-occurrences that I am proposing are distinct from two other constructions often found in Natural Language Processing research:

- Bigrams are often two contiguous words. The words in a co-occurrence can be separated from each other by as many as $n-1$ words, where n is the number of words in the clause; at the same time two contiguous words are not co-occurrences if they belong to different clauses. Where bigrams can be located by simply scanning the tokens in a text, co-occurrences require the system to parse the texts first.

- Collocations are normally thought of as pairs of words that, like co-occurrences, are found near one another a statistically significant number of times. They can be two words that express one thing or one thought – such as "absolute zero" – two words that express a precise notion in physics, or they can be common pairings of words that are not as closely bound as the first kind. For example, "fine dining" or "blue skies". These have an idiomatic flavor. Co-occurrences, in contrast, are not related to any distribution of words in a large sample of language, and can be a unique event.

There are several important issues that are addressed in the following sections.

- How will the clause structures be recognized – what tools will be sufficient? Below the choice of a full probabilistic parser will be discussed.

- Will it be necessary to identify and label all the arguments in a verb structure?
- Will nonfinite structures be recognized as atomic entities or remain embedded in a parent clause? Figure 4.1 contains an example of this, where the text reads, *following the fatal Concorde crash near Paris last month.* Under the Treebank style of annotation, this node is labeled a prepositional phrase (PP) and is not a separate sentential S node.
- What words will be used to form co-occurrences – for example, heads of phrases only or all words; will adjectives be included?
- What kind, if any, of reference and ellipsis resolution will be necessary?

In the rest of this chapter, Section 4.1 discusses the shortcomings of sentence-based approaches involving sentence-by-sentence comparisons; Section 4.2 discusses the choice of the full parser and how it handles syntactic structures; Section 4.3 looks at reference resolution problems; Section 4.4 details the different ways the co-occurrence are selected.

4.1 Sentences

4.1.1 Sentence Structure

Sentences are a commonly used unit of text in Information Retrieval and Summarization, and they offer the advantage of being relatively easy to recognize so that documents can be partitioned into sentences quickly. In English, they start with a capital letter and end with a period, but sentences can be arbitrarily complex, encompassing a number of atomic statements, and they can be broken up for reasons of style and rhythm, rather than content. In other words, the amount of information that any sentence can contain can vary considerably. Although a reader, including a computer program, can count on the coherence of sentences, there are no rules of what can and cannot be glued together.

Figure 4.2 provides an indication of how differently a single key fact can be couched as reporters follow a news event. The event is the crash of the Concorde jetliner on July 25, 2000. These nine selections come from 25 documents published by only the three sources

used in the TREC 2004 Novelty Track. They are not nearly the complete reporting by these three sources on the event. The bold face chunks contain the single fact that the French suspended Concorde flights. The nine variations are, of course, pulled out of context and each fits into the article fluently, as these are professionally written texts. The variations are largely syntactic – the verb is either *grounded* or *suspended* – and the words *flight* and *plane* are used synonymously.

There are many possible reasons for the variety. For one thing, the news organizations are in competition to present different reports even when events have remained the same, and they try to avoid direct repetition even across days. Variety in prose is, in fact, standard advice to students.

In the classic book on written English, Strunk [Str18] notes that a succession of complex (or, in his terms, "loose") sentences "becomes monotonous and tedious." He refers to qualities such as "mechanical symmetry" and "sing-song" rhythm. The remedy is to recast enough of the long sentences "to remove the monotony, replacing them by simple sentences, by sentences of two clauses joined by a semicolon, by periodic sentences of two clauses, by sentences, loose or periodic, of three clauses."

A website offering a guide to writing term papers[2] recommends that students "vary your sentence structure." Under the head of "Combine short sentences", the guide says it is important for good writing to be appealing for the ear: "Try reading your paper out loud. If it seems choppy it can likely be remedied by your grouping short sentences into longer, more complex ones."

These are typical of standard advice given to students for producing clear and effective prose. In journalism, the layout of newspapers and time constraints of broadcasters also require rearrangements of the content. On the surface, one might imagine that relative clauses contain subordinate or parenthetical information, but removing a complete sentence and embedding the information in another sentence can often save a line or two of type, making the article "fit"[3] This discussion about writing style suggests that the content of sentences vary for many reasons, producing structures that are not easily comparable. The

[2]http://writing.englishclub.com/tpe/l8.html.

[3]I worked in the newspaper business for years and am reporting personal knowledge.

1. Xinhua, July 26: Answering questions on the French television, he said that the commercial **flight of Concorde will be suspended** until all guarantees are ensured and the black boxes of the crashed Concorde are studied

2. Xinhua, Aug. 1: **Air France suspended all Concorde flights** after one of its planes crashed last Tuesday in Paris, killing 114 people, 96 of them German tourists.

3. Associated Press, Aug. 2: **Air France grounded its remaining five Concordes** immediately after the crash.

4. Associated Press, Aug. 2: On Tuesday, French authorities decided to keep **its fleet of Concordes grounded** while investigators try to recreate the events that caused the crash near Paris.

5. Associated Press, Aug. 3: **Air France Concorde flights were immediately suspended.**

6. Xinhua, Aug. 4: **Air France has grounded all of its Concorde planes** after the tragic crash in which 114 people were killed, 96 of them German tourists.

7. Xinhua, Aug. 8: The French government, which **suspended the flight of five other Concorde planes of Air France** in the wake of the crash, has come under pressure from some experts and pilots calling for a resumption of the flight of the planes.

8. Xinhua, Aug. 11: France said on Friday that **it will maintain the suspension of the flight of Concorde supersonic aircraft** because the cause for the crash of the Air France Concorde on July 25 has not been definitely determined.

9. New York Times, Aug. 15: When it happened, **Air France immediately suspended flights by its five remaining Concordes**, but, after an initial suspension, British Airways continued to fly the needle-nosed airliner, which crosses the Atlantic in three and a half hours.

Figure 4.2: Variations on the way France's suspension of Concorde flights appeared in the weeks following an air crash on July 25, 2000. The examples were taken from the TREC Novelty Track in 2004.

next section provides further evidence of this point.

4.1.2 TREC Experience

The TREC Novelty Track provides experimental results on the utility of making sentence by sentence comparisons in determining novelty on the sentence level. I'll discuss my results at TREC 2004 in detail in Chapter 6, but in both 2003 and 2004, most groups used sentence comparisons, and in both years a similar pattern of results emerged: virtually no sentence technique was able to achieve precision much better than guessing.

At TREC in 2003, a total of 11 groups participated in Task 2 of the Novelty Track. This task focused on the novelty judgments in isolation of the relevance judgments, and participants were given the relevant sentences selected by the human assessors. Each group could submit as many as five runs. Almost all of the groups performed sentence-similarity calculations, using a variety of techniques, including singular value decomposition, graph-matching and word overlap, as described in Chapter 2. One of the participants counted previously unseen words.

The striking feature about results was the narrow range that was achieved, particularly in terms of precision. All but one of the groups submitted runs that achieved more than the level that would be expected by simply guessing the majority class, a strategy which would yield a precision of 0.65, but for most of them the improvement was relatively small. In all, there were 44 runs submitted. Table 4.2 shows the distribution of precision scores. Of the total, 34 were between 0.65 and 0.74.

On the other hand, Table 4.3 shows that recall was relatively high for most submissions, and was distributed much more evenly from the level of 0.65 and 1.00. Half the systems obtained recall scores of 0.85 or more, and of these, 12 systems obtained scores of 0.95 or better.

These distributions are interesting in that they suggest it is much harder to be accurate than it is to be inclusive. When the systems are ranked by an F-measure equally weighting precision and recall[4], the ranking favors those that return larger numbers of sentences,

[4]F-measure is a combination of precision and recall and the computation to give each equal weight is $F = \frac{2PR}{P+R}$.

Range of Precision Scores	Number of Submissions in Range
Less than 0.65	5
Between 0.65 and 0.69	13
Between 0.7 and 0.74	21
More than or equal to 0.75	5

Table 4.2: The distribution of precision scores among the 44 submissions to Task 2 of the Novelty Track. About 65% of the sentences marked as relevant by the human assessors were found to be novel.

---

meaning that there will tend to be less compression, and at the same time some information loss.

In a real world system, such a system would fail to satisfy users. The system cannot provide high precision for those who value accuracy, and it would not be of much benefit to those who value recall since the systems obtain their high recall by giving the user most of the original inputs. In the later case, the user could forgo the system and just take the original texts.

When the scores are grouped by participants, the results are interesting. In some cases, different parameters and different strategies showed no difference in the final numbers. IRIT (Institut de Recherche en Informatique de Toulouse) was such a case, where all four runs submitted scored an identical 0.71 in precision and 0.76 in recall (the group's fifth run was a baseline of accept everything, which of course got a precision of 0.65 and a recall of 1.00, and a higher F-measure than its other runs.)

In the classic retrieval task, as the systems become stricter on what units (no matter whether documents or sentences) are selected, precision goes up and recall goes down. In the novelty track, a change of strategies often showed both precision and recall moving in tandem, often in big jumps. Take, for example, the case of the Institute of Computing Technology at the Chinese Academy of Sciences(CAS-ICT). Four of its submissions scored a precision of 0.65, and a recall of either 0.73 or 0.74. The group used Maximal Marginal Relevance, developed at CMU, [GC98] for three of these submissions and a fourth system

Range of Recall Scores	Number of Systems in Range
Less than 0.55	4
Between 0.55 and 0.64	1
Between 0.65 and 0.74	7
Between 0.75 and 0.84	10
Between 0.85 and 0.94	10
More than or equal to 0.95	12

Table 4.3: The distribution of recall scores among the 44 submissions to Task 2 of the Novelty Track. About 65% of the sentences marked as relevant by the human assessors were found to be novel.

---

to set the number of selections dynamically. The fifth run "only applied the simplest techniques, i.e. comparing the words that occur in both sentences", but it achieved a big boost in performance, scoring a precision of 0.73 and recall of 0.87.

Like the TREC participants, I also implemented a relatively simple sentence-based system and applied it to my own data and the Novelty Track data from 2003 and 2004. The system is a unweighted vector-space model using cosine distance as the similarity metric. As the system scans the input sentences, similarity is computed against each previous sentence. If the maximum similarity score exceeds a threshold, the sentence is labeled old; if it doesn't, the sentence is labeled novel.

$$Novel(s_i) \begin{cases} true & \text{if} \quad Cos(s_i, s_j) < T, \quad \text{for} \quad j = 1 \ldots i-1 \\ false & \text{otherwise} \end{cases}$$

I tested this system with thresholds ranging from 0.10 to 0.90. When the cosine threshold is set at the low end, the system finds that most sentences are similar to some previous sentence, and classified as not novel. This setting should be the most stringent for novelty, making sure there is no overlap whatsoever between the new and old sentences. Yet, as Figure 4.3 shows, there are many errors, and precision reaches a maximum of < 0.8, actually beginning to turn downward. This result suggests a need for a way to find semantic equivalence, a topic we will address in Chapter 5.

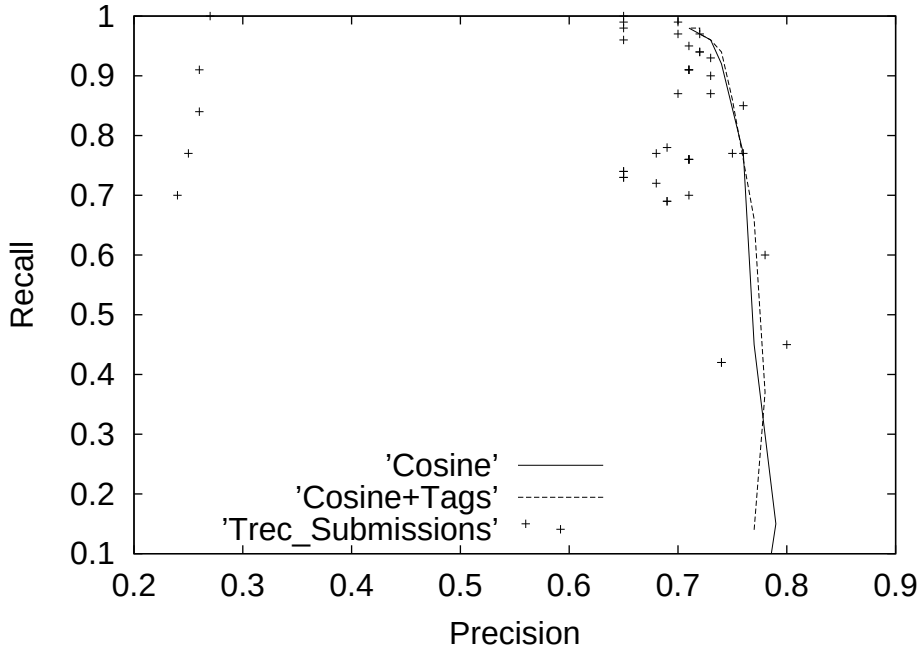

Figure 4.3: The scatterplot shows the performance of all the submissions at TREC 2003. The solid and dotted lines show the performance of two variations of a baseline vector-space system.

When the cosine threshold is set at the high end of the scale, very few sentences are disqualified as too similar. Figure 4.3 shows that even at the middle level, the system returns most of the input sentences as novel.

Figure 4.3 shows the combined results at the Novelty Track compared with the performance of my simple vector-space model. Like CAS-ICT system, our simple, straightforward approach did as well as the top systems, but the problem is that the graph suggests a ceiling on the results. The experience of the participants in the Novelty Track represents a wide variety of sentence-based strategies and gives no strong indication of how to improve the results. I am suggesting, simply, that this experiment provides strong evidence that sentences are not a good unit to work with. Since sentences can be composed of two or more S-structures, some of which may be novel and others not, there is no satisfactory way to average the content.

4.2 Parsing

In broad terms, my goal in parsing is to decompose the sentences into units that are as close to indivisible as reasonably possible. Each unit should express some relationship between entities, and most of the units should express only a limited number of such relationships. The problem I described in section 4.1 is that a large number of sentences are often constructed of several sentential structures. Any number of grammatical formalisms would probably be acceptable so long as they can partition sentences into smaller chunks accurately. In that way, the entities in a text can be linked to other appropriate entities, without superfluous, noisy links being established. A study of grammars and linguistic theories are beyond the scope of this thesis; I am using Charniak's probabilistic parser [Cha00] which reproduces the grammar in the style of the Penn Treebank [MSM93]. It is one of a number of probabilistic parsers trained on the Treebank, all of which are fairly accurate, especially on recognizing the large chunks we are seeking. In addition, they provide extra power to accurately recognize many constituent phrases. But they are also relatively inefficient, both in terms of speed and resource use.

I should make clear at this point that I am not using the full resources of the parser, but want to narrow down the size of the units to limit the formation of co-occurrences to pairs of words that truly interact with each other. In this way, I hope to make sufficiently reliable estimates of a long list of binary relationships in two documents that can be compared, and points of difference – i.e. novelty – can be established.

Figure 4.4 shows a text version of a parse tree of another sentence in the Concorde cluster.

The shading shows the clauses that are separated from one another. The parser provides the part of speech tags and structure to the phrases. The part-of-speech tags allow the program to choose the content words that are desired and leave out the function words in forming the co-occurrences, which are described in more detail in Section 4.4. The phrase structure provides enough information to easily select the head words, and if desired, to choose the major constituents of each clause.

The goal is to get a list of co-occurrences, which are pairs of words that are syntactically linked to one another, such as those in Figure 4.5. Equally important is avoiding

```
(S1 (S (NP (DT The) (JJ daily) (NNP Liberation))
    (VP (VBD said)
      (SBAR (IN that)
        (S (S (PP (RB ever) (IN since) (NP (PRPS its) (NN conception)))
          (, ,)
          (NP (NNP Concorde))
          (VP (VBZ has)
            (VP (VBN lost)
              (PP (TO to)
                (NP (NP (JJ market-oriented) (NNS planes))
                  (PP (IN like) (NP (NNP Boeing)
                                    (CC and) (NNP Airbus))))))))
        (, ,)
        (CC but)
        (S (NP (PRP it))
          (VP (VBZ remains)
            (NP (NP (DT the) (JJS best) (NN plane))
              (PP (IN in) (NP (DT the) (NN world)))))))))
    (. .)))
```

Figure 4.4: A parsed sentence with major clauses shown in shaded blocks.

inappropriate links across clause boundaries. The newspaper *Libération* is not a loser in the marketplace, and those words do not form a co-occurrence.

The parsed sentence makes apparent two other issues that must be dealt with:

- What words are allowed to form co-occurrences. If co-occurrences are formed for all content words in a clause, then a connection is drawn between *conception* and *Boeing* in the example above. Four strategies were tested:
  - Use all words. In this case, the selections are always as accurate as the underlying parse, but some noise would be expected, especially in larger clauses.
  - Use all nouns and verbs, ignoring adjectives. The motivation is to try to focus on the more important terms and avoid noise, but in some cases, the adjectives supply a considerable amount of the content, as in the phrases “market-oriented planes” and “best plane” in the example in Figure 4.4.

$Liberation \Longleftrightarrow say$

$Concorde \Longleftrightarrow lose$

$lose \Longleftrightarrow Boeing$

$lose \Longleftrightarrow Airbus$

$Concorde \Longleftrightarrow Boeing$

$Concorde(it) \Longleftrightarrow bestplane$

Figure 4.5: Examples of co-occurrences that are to be extracted from the parsed sentence in Figure 4.4.

  - Use only the head words, or most important words in each phrase. In some cases, using head words would be similar to ignoring adjectives, but in other passages, noun phrase content varies more. Several methods of importance were tested and will be discussed in Chapter 5.
  - Use co-occurrences from only the major constituents in a clause, the subject, object and verb; from these three constituents, all three possible unordered pairs are taken. Then, for each of the three, additional co-occurrences are collected, combining head words with modifiers. In this hierarchical setup, in the middle clause (unshaded) in Figure 4.4, co-occurrences would first be formed for "Concorde" and "lost", and "Concorde" and "planes", and "lost" and "planes". Then, in the next stage, more co-occurrences would be formed from "planes" and "Boeing", and "planes" and "Airbus."

- What effort will be made to resolve references? Reference resolution will be discussed further in Section 4.3, but in Figure 4.4, in the lighted shaded clause at the end, *Concorde* would be ideally be substituted for *it*.

By employing a powerful probabilistic parser, the program has better information for building its internal representation of both new and old inputs. Early in the work for this thesis, I used a fast, finite state parser that does a reasonable job locating clause boundaries.

This parser, called Clausit, was initially developed for joint work with Mitre [SMC01] for a biographical summarizer. It uses a cascade of finite state machines implemented by

CASS [Abn96], and tries to find clause boundaries. It pieces together syntactic patterns that do not necessarily correspond to distinct noun phrases – like the clause chunks labeled *XL*, or *XS*. Example output from Clausit in Figure 4.6 is quite close to the probabilistic parse. The main error concerns the prepositional phrase between the first and second clauses. We chose to switch to a probabilistic parser in order to get more accurate clause boundaries and at the same to have a good representation of the constituents that make up a clause. In addition, the probabilistic parsers produce more reliable part-of-speech tagging than standalone taggers. In fact, the Charniak parser dispenses with pretagged input and exclusively supplies its own tags, while others, like the Collins parser [Col96] requires tagging, but freely changes the tags of those words for which it has probabilities.

To complete the structural representation of the text, we use a named-entity recognizer, Talent [RWC97], which is applied to a concatenation of all the input texts in a cluster, achieving an approximate cross-document coreference for named entities. Thus, the initial text processing comprises parsing and named-entity recognition done in parallel and the results are merged. This operation is made somewhat easier because we mark the sentence boundaries before either Talent or the parser is started, guaranteeing that the sentences always contain the same content. In merging, named entities are governed by Talent's markup and all other part-of-speech tags are taken from the parser, as are all structural annotations.

4.3 Coreference

Figure 4.4 suggests that accurate coreference resolution could add considerable power to the system. In the example, the key word *Concorde* is lost in the last clause, where the aircraft is referred to as the "best plane in the world." The example is local in that the anaphor, the pronoun *it*, is in the same sentence as its antecedent. The clause-based approach in this thesis draws attention to the general problem of reference resolution, but sentence-based systems face the issue, too. A full exploration of reference resolution is beyond the scope of this work, but still it is an issue that I have tried to address without encouraging results, achieving less than 50% accuracy on subject and object case pronouns in preliminary tests.

```
<c>
xl::(dt,The,The) (jj,daily,daily) (name,Liberation,Liberation)
(vbd,said,say)
comp::(comp,that,that)
rb::(rb,ever,ever)
pp::(p-comp,since,since) (prps,its,its) (nn,conception,conception)
cma::(cma,,,,)
</c>
<c>
xl::(name,Concorde,Concorde) (v-has,has,have) (vbn,lost,lose)
pp::(to,to,to) (jj,market-oriented,market-oriented) (nns,planes,plane)
cs::(cs,like,like)
np::(name,Boeing,Boeing)
cc::(cc,and,and)
np::(name,Airbus,Airbus)
cma::(cma,,,,)
cc::(cc,but,but)
</c>
<c>
xl::(prp,it,it) (vbz,remains,remain)
np::(dt,the,the) (jjs,best,good) (nn,plane,plane)
pp::(in,in,in) (dt,the,the) (nn,world,world)
.::(.,.,.)
</c>
```

Figure 4.6: The alternative parser that offered greater efficiency but is not as reliable in recognizing the hierarchical structures

A fair amount of study has gone into pronoun resolution, like the example here, where *it* refers to the *Concorde*. The obvious strategy would be to borrow an already developed system. In pronomial resolution, researchers have reported accuracy of 70% or better, but when several systems that are available were tried on randomly selected articles, they often fell far short of that accuracy. In addition, the algorithm for new information detection presented here would also benefit from the resolution of various kinds of ellipses. Thus I decided to incorporate some facility for handling pronouns and ellipses in the program, allowing the system to be run with or without these functions. I will leave the discussion of deeper references, various kinds of definite references to both named entities and common nouns, as well as predicate references, to Chapter 5.

The resolution strategy operates mainly on surface information. A focus stack is built as each article is scanned, containing pointers to the program's internal representation of each noun. The system uses such features as syntactic role, number agreement and distance to order the candidates and select the first one that fits. The parser does not identify the syntactic roles, so I rely on matching a series of regular expressions to determine the subject, object and main verb of each clause unit. Where the subject is missing, for example in cases of nonfinite verb phrases, I take the nearest noun-phrase head.

By implementing coreference functionality within the program, I also avoid the problems of weaving together the output of separate, existing systems.

To test **NIA**'s resolution system, I compared the result of the two articles in the Concorde crash set that had the largest number of pronomial references against two systems, one by Siddharthan [Sid03] and the other by Morton [Mor00]. These were both developed recently and are representative of available pronomial-resolution systems.

Table 4.4 shows that there is little difference among the performance of the resolution algorithms. The **NIA** system and Siddharthan's are rule based while Morton's uses statistical models. All three use WordNet to some extent to filter plausible antecedents, namely whether the pronoun is matching a person or a thing.

The articles were chosen from among the 25 on the Concorde crash that are being used as examples in this chapter, but they are the two with the largest number of personal pronouns. The distribution of pronouns in the 25 articles is interesting. The two with the

Article	Total Count	Number Correct		
		NIA	SID	MOR
APW	16	3	4	3
NYT	14	8	7	6
Total	30	11	11	9

Table 4.4: Results of three systems on personal pronouns in two articles on the crash of the Concorde jetliner in 2000 in Paris. The systems are the internal algorithm in **NIA**, and the systems by Siddarthan and by Morton. Only personal pronouns were examined. **NIA** does not operate on possessive pronouns, since these tend to be local and are often covered within a clause, leaving nothing for the system to gain. Siddharthan does personal, possessive and relative pronouns, while Morton does personal and possessive pronouns.

---

most pronouns are, as expected, the largest, at 750 and 683 words, with the average article at around 275 words. Four of the articles have no personal pronouns, and nine have only one. The rest of the articles have between 2 and 6. The performance of the resolution systems and the relatively low number of pronouns in these news articles makes it difficult to gauge the benefits. The results in Chapters 7 and 8 are better when pronoun resolution is turned off.

One thing is clear: The difficulty of pronoun resolution varies widely. There are easy ones, such as this from the New York Times article:

> ... after **investigators** said **they** would urge the withdrawal of the airplane's airworthiness certification following the fatal Concorde crash near Paris last month.

All three of the systems got that right. But some passages are quite difficult, requiring a fair amount of semantic and pragmatic information, like this one from the Associated Press:

> Yet within hours of the crash, the French began mourning the death of the Concorde, a mix of technology and elegance that was the pride of this Gallic

nation.

"**It (the Concorde)** was 31 years old.

For France, **it (pleonastic)** is a day of mourning.

... **It (the Concorde)** will remain the myth of the beautiful white bird,"

Le Figaro said .

That was beyond the reach of any of the three systems.

4.4 Extracting Co-occurrences

Once the input documents are parsed, the system extracts 24 features for each unit (i.e. the clauses) including the co-occurrences, structures that hold pairs of words assumed to be linked in some relationship to one another. The system assumes that all words in a clause are in some way related to one another. The full complement of features will be discussed in Chapter 7.

The overall idea is to compare what entities of document d_n are covered by documents $d_1 ... d_{n-1}$. What is not covered is therefore *novel*. The entities can be named or not, abstract or not, and actions or objects – in short anything realized by a content word. The system makes the coverage judgment on the basis of surface information. An interaction or relationship between two entities e_i and e_j is *not covered* if it exists in the current document, d_{curr} but not in the background bg. A relationship is defined as existing between two entities if two words referring to those entities occur in some clause c.

$$
\begin{aligned}
&Nov(e_i, e_j) = True, \\
&If R(e_i, e_j) \in d_{curr}, R(e_i, e_j) \notin bg, \\
&Where R(a, b) \rightarrow a \in C_c, b \in C_c, \\
&And C_c is any clause in either d_{curr} or bg
\end{aligned}
$$

To compare documents, the system compares the entities in one document with their equivalents in previous documents in the *background*, and determines efficiently which are covered by the previous documents and which are not. The complexity of the syntactic phase is dominated by the parsing process. After parsing, the co-occurrences are read directly and

compared by checking if both elements match. In the simplest case, the current document may have a previously unseen entity, which in itself would be evidence of novelty. It is clear that a clause containing the name of an entity – a person, place or thing or idea that had not been mentioned before – contributions some number of novel entities. In more complicated cases, the new document may not mention any new entities, but contain many new co-occurrences.

No effort is made to specify the semantic roles involved in relationships between words. In many cases, the system relies on the fact that news events are usually consistent, and that an argument of one relation doesn't switch roles from one article to the next – i.e. the victim of one article doesn't often suddenly become the criminal in another. In a way, this research is an exploration of how little understanding we can get away with, assuming, of course, that the input documents are grouped together by a reliable clustering algorithm. Further, only one sense of polysemous words is assumed to appear in one set of articles, borrowing and expanding the principle of "one word sense per discourse" [GCY92]. Likewise, I left for future work the difficult problem of negation, which would require the ability to determine the scope of the negation. Early in the research, I searched for examples of statements and their direct negations, but had a great deal of difficulty in locating these in the news domain. Often the negation was expressed with much different terminology.

The procedure decomposes a document into structures that make it easy to compare to the previous documents. By doing this transformation, the system avoids pairwise similarity judgments of syntactic units, such as sentences or clauses. Instead, a history is cumulatively built up to hold all of the entities and co-occurrences in the documents. As each document is processed, this history, or *background*, is updated.

In the Concorde set of articles, Figure 4.7 shows the complete short article appearing soon after the crash. Soon after it appeared a somewhat longer article with more details appeared.

The first paragraph of the later article read:

> PARIS,July 25 (Xinhua) – A Concorde supersonic passenger plane of Air France crashed Tuesday afternoon soon after taking off from the Roissy international airport north of Paris, **killing all 109 passengers and crew on board and**

PARIS,July 25 (Xinhua) – One survivor, possibly among the 109 passengers and crew members of a Concorde passenger plane of Air France which crashed Tuesday afternoon near Paris, was found, said rescuers.
But the survivor could also be someone on the ground when the crash took place in what is one of the biggest air disaster in 30 years.
The plane came down on a hotel and its restaurants, possibly causing more casualties.
All the passengers were Germans going to New York with a tourist company, where the crew members were French.
The cause of the crash is unknown yet, but a witness said that he saw the left engine of the plane was on fire and crashed two minutes later.

Figure 4.7: One of the first articles on the crash of the Concorde in July 2005.

---

four people on the ground, said the French Interior Ministry and rescuers.

The verb *kill*, which does not appear anywhere in the first article, forms co-occurrences with each of "passengers", "crew", "board", "people" and "ground". Later in the article, "kill" appears again:

> The four people **killed on the ground could be in a hotel near the airport** on which the plane came down.

In both sentences, the noun *ground*, which is in the earlier article, also forms novel co-occurrences with *hotel* and *airport*, also in the earlier article, amplifying that there is a new arrangement of words.

I do not argue that each of these co-occurrences by themselves makes for novelty, but they are evidence for inferring what is novel. The system weighs each of them, according to their frequency in the set and, optionally, according to a semantic-content weighting that will be discussed in the next chapter.

The co-occurrences drawn from the syntactic analysis of the input documents are the heart of the *Micro View* of the problem. The *Macro View* will be discussed at length in Chapter 6 on contextual issues. The following chapter will discuss the semantic analysis, which expands the *Micro View* processing from one based on words to one based on concepts.

4.5 Conclusion

In this chapter, I have shown that sentences are a poor choice for a task of this nature. Sentences are often complex structures that can decomposed into a number of clauses. These clauses can be reassembled by different writers without any change in meaning. Their motives are varied – sometimes seeking to achieve a particular artistic or emotional effect, or to change the emphasis of article, or even to simply make the article appear different from those produced by competitors. By using sentences as the unit, a system must make similarity, and consequently, dissimilarity judgments, on the basis of arbitrary choices by the writers. The strategy to overcome the problem with using sentences is to decompose the sentences into clauses, and then to extract co-occurrences from the clauses, i.e. pairs of co-occurring content words. Chapters 7 and 8 will show that these co-occurrences provide powerful features to be used by learning algorithms to identify novel passages.

Chapter 5

Semantics

Semantics, or some approximation of it, is necessary for the new-information task. Facts and events can be expressed in many different ways, and a new-information system must have a means to determine which expressions are equivalent. My task, at its most fundamental level, is to establish whether or not two atomic expressions in text are equivalent. If I could reasonably expect to answer this by string matching, I would have no need to delve into semantics. But since there are ships and boats, rebels and fighters, scholars and researchers, string matching will not take us very far. We need a practical approach to the subject of semantics.

Researchers in Information Retrieval can successfully treat documents as bags of words for such tasks as searching and document clustering. There are usually enough words in a document to distill or approximate the document topic. When the task is classifying at sentences or even paragraphs, the ability to approximate is lost. Single statements are are often incomplete. For example, consider the statement, "This is a big victory for us." By itself, without context, this sentence cannot be understood, yet it is clear that it conveys an important fact for whatever narrative it belongs to. In fact, it was easy to find this sentence in a large number of documents on the Web, concerning such diverse victors as the Miami Dolphins, Oracle in its competition with Microsoft, and video buyers of the movie *Babylon 5*. Another detail won't necessarily help pin down this variation: "The Supreme Court decision is a big, big, big victory for us." The Supreme Court is often concerned with technology and consumers, and occasionally with sports, but in this case, the decision is

about affirmative action (from the perspective of a civil rights group.) The difficulty is even greater across documents than within documents. In the cross-document case, different writers may have different perspective, knowledge and taste in word choice. The word victory could easily be "triumph", or "success", or "win".

In its broadest sense, the semantics needed for new-information detection would provide the ability to equate disparate references to entities or events. Named entities are one subproblem where some tools exist. Named-entity recognizers identify different forms for the proper names of entities, whether, persons, places or things. I use a version of *Talent*[RWC97], a named-entity recognizer from IBM on the concatenated documents in order to achieve some cross-document normalization of names; this stretches the capability of the tool somewhat. The available named-entity systems are outgrowths of the Message Understanding Conferences(MUC) in the 1990s, which addressed information extraction issues in a domain-specific situation. As such, they were concerned with an equivalence problem: how to match statements in texts to specified slots in a template. For example, one of the early MUCs used terrorism as the subject. The template slots were such items as terrorist groups, terrorist attacks, casualties, etc. At the last MUCs, evaluation was divided into several tasks. The best results were for named-entity recognition, on which many systems had F-scores of 0.9 or better, but the most difficult was the Scenario Template task, which required participants to relate entities to actions and to one another. The templates contained a number of slots for elements typical of the event under consideration – "bomb" and "victim" in the terrorism exercise, for example. The best S.T. F-scores were around 0.5. Since 1999, the Government has organized a kind of successor to the MUCs, the Automatic Content Extraction(ACE) evaluations, which are more general, i.e. domain-independent, seeking to link all references to entities. This is the kind of capability that the new-information task requires, linking all references to an entity, whether realized as proper names, synonyms, pronouns or any form of anaphora, across a number of documents. But this is hard. As Professor Grishman at New York University [1] puts it:

[1]Professor Grishman has been involved in both the MUC and ACE evaluations. His discussion here is taken from lecture notes for his course, *Advanced Natural Language Processing*, at NYU in the spring of 2004. They can be found on the web at http://www.cs.nyu.edu/courses/spring04/G22.2591-001/. The quote was

> "Task definition and data preparation is much harder for free text extraction because it involves the interpretation of the information conveyed in text – information which can be described in many different ways. ...
>
> "In the ACE evaluations, the Government has shifted to more general relations and events, such as a person is at a location, a person has some social relation to another person, etc. Unfortunately, these relations have been considerably harder to pin down than the MUC events."

This situation presents a dilemma. The chosen task, new-information detection, requires powerful tools to make use of deep semantics, but these are not available. In an overview of the IBM system for the ACE evaluation [ILK+03], researchers report on their composite system, which obtained a F-measure of 0.74 in the September 2002 in the "mention detection" task, where all types of references to named entities are to be identified. The authors describe this as competitive at the evaluation. This score includes the identification of the explicit named entities, which were handled by the MUC-inspired tools like Talent. These generally obtained F-measures in excess of 0.90, so that the harder instances, such as definite noun-phrase references and pronouns, deteriorated accuracy substantially. There has been work in semantic parsing in which the arguments to verb forms are identified, but performance is not yet strong. Gildea and Jurafsky [GJ02] report 0.65 precision and 0.61 recall in working on example sentences from Framenet [BFL98], a manually labeled corpus of verb frames. The construction of a semantic dictionary would mean a whole thesis in itself. Shallow semantics, however, could prove helpful. This chapter will deal with using, combining and even extending some existing resources to obtain the necessary surface semantics, by which I mean using some method of word similarity to determine the similarity of statements across documents.

In the next section, I will talk in more detail about the general problem, and then examine some of the problems with WordNet as it exists in Section 5.2. I will describe the lexicon in more detail and discuss experiments with adding information collected by large-scale statistical analysis of corpora in Section 5.3.

---

taken from lecture 11.

A novel feature of my work concerns efforts to identify words that carry a large amount of content, in contrast to very vague and general words. Linguists talk of "empty nouns," like "someone," "someplace." I want to make additional distinctions in order to treat very general words, like "idea" or "choice" that do not suggest any particular topic in the news genre. My use of co-occurrences described in the previous chapter is prone to false positives when such low content words are linked to other words and compared to co-occurrences from other documents. To a person, or a hypothetical system that could understand deep meaning, these words are very important, but my notion of co-occurrences operates only on the surface. I will discuss our efforts to identify these words and use them in Section 5.5.

Next, I will discuss the problems in lexical semantics further, and then look at WordNet's strengths and weakness for addressing those problems. Then I will describe testing of an automatically obtained thesaurus and efforts to combine it with WordNet.

5.1 Wealth of Words

Writers place a premium on inventiveness and creativity, on finding new ways to evoke emotion, on using words in new and different combinations. Word play, puns, unusual metaphors are all a basic part of humor. These values are not only found in poetry and literature, but also in more ordinary venues. A cursory search on Google for "puns" or "wordplay" turns up hundreds of web sites devoted to this form of amusement. But puns and metaphors that are overused become clichés, which experts frown upon. Variation is a constant theme in advice on good writing. A web-based writers' handbook for students [2] urges young writers to vary word choices, as well as sentence structure.

> While repetition can be used to achieve coherence, it must be used sparingly. Use synonyms whenever possible to avoid using the same word over and over again in your writing. Also, it is important to vary your sentences. Look at the sentences in each paragraph of your essay. How many begin with the subject? How many begin with the word "the," or a noun, or a name?

[2] http://www.kent.k12.wa.us/KSD/KR/WRITE/FIVE/revise_style.html

Slang and jargon are nothing more than variations of literary practices. What's worse, much discourse is aimed at persuasion or obfuscation. To listen to a political speech, or to attend a business meeting, is like stumbling upon an Escher drawing (like the one in Figure 5.1), where it's hard to know up from down, or right from left, coming from going. The utterances will have an emotional flavor that sounds like one thing, but carry enough ambiguity to mean another.

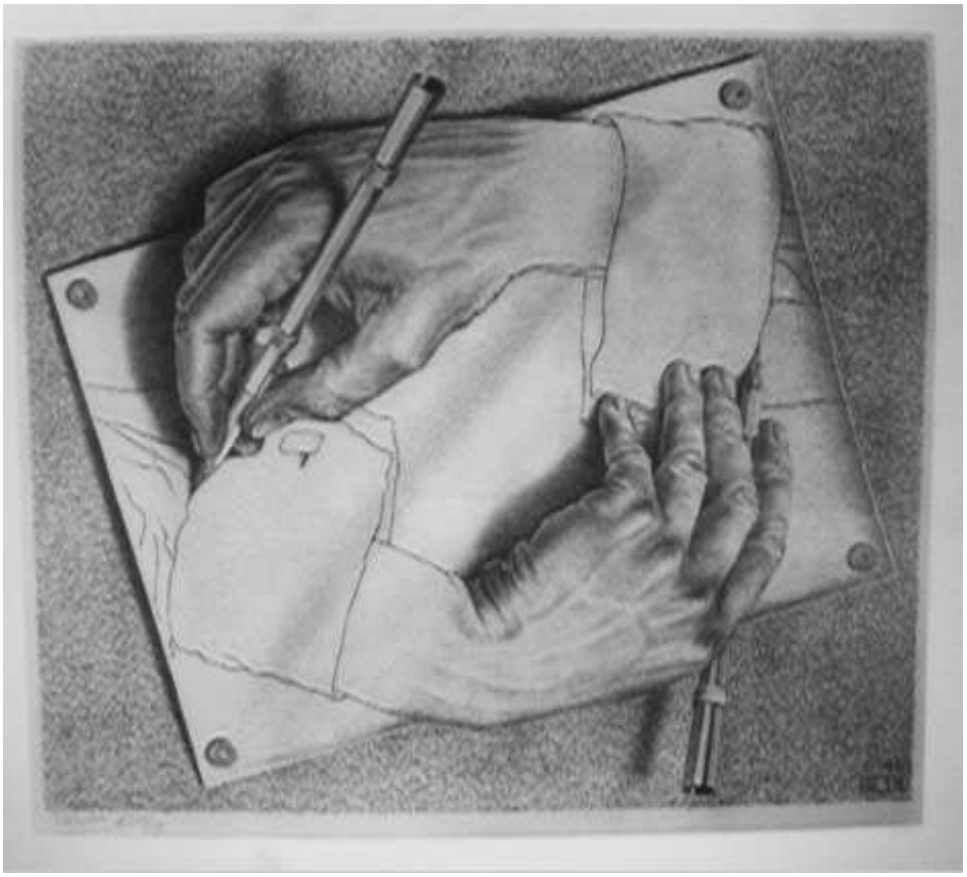

Figure 5.1: Escher's Hands

Although news professionals seek to restrain the level of ambiguity in their reports, they are under pressure to find information that their competitors do not have, and to produce the next day's reports on hot topics even when the events don't materialize. Broadcasting, and recently the World Wide Wide, has intensified the demands on the media, with a growing number of 24-hour news outlets. The journalistic solution is often to fall back on

the human facility to produce language made up of shadings and colorations, hints and feints.

Lakoff [LJ80] argues that much of human communication is conducted with metaphor, where words are not quite what they seem on the surface. Dictionaries do not help. Jurafsky and Martin [JM00] provide this example from the American Heritage Dictionary [Mor83]:

right located nearer the right hand esp. being on the right when facing the same direction as the observer.

left located nearer to this side of the body than the right.

red the color of blood or a ruby.

blood the red liquid that circulates in the heart, arteries and veins of animals.

It's easy to see that machine readable dictionaries won't help much in the new-information task without a human's knowledge of the world. The entries for "right" and "left" are self-referential, and the entries for "red" and "blood" are circular as to color. The entries for "right" and "left", in addition, are not in any consistent format. The entries for "red" and "blood" are better in that "red" is identified as a color and "blood" as a liquid, but think of the associations that "blood" conjures up in our minds, and how impoverished the definition is. Most utterences that contain the word "blood" would be incomprehensible with the given definition. Every text, even a professionally written, straightforward news article, exists in a world that is familiar to human beings but that is not encoded in any way a machine can use. I raise this issue of the complexity of human language to recognize the limitations of available tools so that system development may proceed with the appropriate amount of caution.

The subject of semantics, or meaning, is vast and this thesis addresses only a very small corner of it; in particular, a small corner of lexical semantics. The discussion does not address the larger problem of language understanding. The goal in this work is to establish whether a particular statement – each atomic statement scanned in turn by the system – is covered by one or more previous statements that the system already scanned and stored in the background structure.

Rather than make these comparisons by entire units, the system looks at the words in pairs, co-occurrences that were described in Chapter 4, and determine whether the particular pairing was encountered before or not. (Of course, there may be 3, 4 or more co-occurrences in any one clause.) In the simplest case, the pair of words comprising one co-occurrence will exactly match the pair in another, therefore providing no evidence of novelty for the current unit, which is a clause in my system. To cope with more complicated cases, I use semantic resources to find additional techniques to make better, more robust comparisons of co-occurrences. I start by using a lexicon, *lex*, much like a thesaurus, to obtain synonyms. Given a word, it will provide other words that can point to the same underlying entity or idea. I'll discuss the lexicon construction in Section 5.2, but here is an example entry for the word *thesis*:

thesis, n, {"dissertation","premise","treatise"}

Once a word is looked up in the lexicon, it and all of its equivalent words are recorded as different ways of expressing the same concept – forming a *concept set*, *cs*. Any subsequent mention of any of these words will be considered a match so that a *thesis* in $document_i$ will be considered the same entity as a *dissertation* later in $document_i$ or in any $document_j$ in the set. The system maintains a dictionary, *dict*, in memory for each set, and adds a new entry to this dictionary each time a new concept is encountered.

```
For each document d
    For each word w
        if w ∈ some dict(cs_i)
            then append w to dict(cs_i)
        otherwise
            fetch the entry for w in lex and create a new dict(cs_i)
```

This strategy stays on the surface, and does not attempt to determine definitively whether an expression, such as a definite noun phrase actually refers to some previous expression, and it does not attempt to establish the thematic roles or to pin down selectional restrictions. It expects that the clustered input sets will be all on topic and thus there is some control on polysemy – because of the "one-sense per discourse" notion [GCY92]. I

began by using WordNet [MBF+90] synonyms in the lexicon, but modified this, as described later in this chapter.

In addition to equating common nouns that reference the same object, the system also uses Talent, the named-entity recognizer as mentioned above, to equate different realizations of proper nouns across the documents in a set, and it has very basic pronoun, and even ellipsis, resolution functions available as options.

As I said in Chapter 4, I rely mostly on structural cues for pronoun and ellipsis resolution, but in future work will try to employ the semantic lexicon to improve performance in this area.

I rely on a named-entity recognizer to equate proper nouns to short versions of their names or to nicknames, but I view the problem of linking proper names to noun phrases headed by common nouns – such as "Columbia University ... the institution" – as an extension of named entity recognition, and that is beyond the scope of this thesis.

Finally, I will also leave paraphrasing for future work. By paraphrasing, I mean the use of different expressions that express the same meaning, often encompassing idiom and metaphor, going beyond strict synonyms. There has been some progress in the area, notably Barzilay [BM01, BL03]. She offers a dictionary of 56,942 paraphrases. Of these, only 6,944 appear more than half a dozen times in the training corpus. Table 5.1 shows 12 examples of the more frequent matches, drawn at random [3]. These are typical of the results. The results are interesting. Some that are correct are simply morphological or syntactic variation, and others share some common traits, but in all they are too noisy to increase accuracy in new-information detection.

But I'm leaving for future work efforts to link proper names to definite noun phrases, including easier cases like "the State Department ... the department" and harder cases like "the State Department ... the government".

[3]Randomized using the built in random number generator of perl, with the system clock as the random seed.

17	became	declared
8	combating	combatting
24	cuts	final agreement
9	jews	militant islamic organizations
13	staff writer bradley graham	times staff writer john hendren
7	had planned	planning
8	bin laden	ubl
9	talk	talks
7	staff writer vernon loeb	wire services
32	new york	world trade center
10	adopting	passing
5	are	re setting

Table 5.1: These 12 entries were randomly selected from Barzilay's list of automatically obtained paraphrases. The selections were limited to those that occurred at least seven times. The first column shows the number of occurrences, the second and third columns are paraphrases of one another.

5.2 WordNet

WordNet [MBF+90] is an extremely attractive resource because it provides a large semantic hierarchy that can be traversed to find other relations in addition to synonomy.

For the work here, WordNet is a natural place to start. Something more than a flat list of synonyms is needed. A new information system has to determine when different words refer to the same underlying object. These are often not synonyms and can be at some distance in the hierarchy, sometimes taking a step in one direction and then turning laterally. For example, a "mugger" is a "robber", and a "robber" is a "thief", and a "thief" is a "criminal". In order to recognize that a "mugger" might be called a "thug", one has to traverse all those steps of hypernymy, and then turn down to a hyponym of "criminal" in order to find "thug".

I am calling such words *Referential Equivalents (REQ's)*, and WordNet captures many of these in its hierarchy of hypernyms, which are more general categories of the word in question, as an automobile is a motor vehicle, and of hyponyms, or more specific types of the word; for example, a *Chevy* is included in the category of automobile. But siblings are difficult, especially if they are not synonyms; for example, automobiles and motorcycles are motor vehicles, but will never refer to one another. This raises the problem of deciding how and how far to look for a *REQ* in WordNet. The list of possibilities can grow quickly in the traversal of the WordNet hierarchy, linking words that would never be linked in text and missing words that are often used as stand-ins.

For an example of the expansion of a word that goes too far, and not far enough, take the noun *cat*. WordNet 2.0 gives us eight fairly distinct senses:

1. cat, true cat ⇒ feline, felid
2. guy, cat, hombre, bozo ⇒ man, adult male
3. cat ⇒ gossip, gossiper, gossipmonger, rumormonger, newsmonger
4. kat, khat, qat, quat, cat, ... ⇒ stimulant, stimulant drug
5. cat-o'-nine-tails, cat ⇒ whip
6. Caterpillar, cat ⇒ tracked vehicle
7. big cat, cat ⇒ feline, felid
8. computerized tomography, computed tomography, CT, computerized axial tomorgraphy, computed axial tomorgraphy, CAT ⇒ X-raying, X-radiation

Sense 1 is clearly the most frequent sense, yet the hyponyms and hypernyms of this sense fail to show us two of the most common words that would refer to a cat:*kitten* and *pet*.

There is only one sense of kitten:

1. kitten, kitty ⇒ young mammal ⇒ young ⇒ animal

And three senses of pet:

1. pet ⇒ animal
2. darling, favorite, pet, dearie, ducky ⇒ lover ⇒ person
3. pet ⇒ irritability
4. positron emission tomography, PET ⇒ imaging

There are no hyponyms for the Sense 1 of pet, and so cat is not an example of a pet. Following the hypernym arcs of cat, Sense 1 is seven steps away from animal. Sense 2 is three steps away from person, and Sense 3 is two steps away from person – an equal distance from the sense of pet in "teacher's pet" to person. This would be a very dangerous path by which to link cat to pet. Starting at some words, the complete subgraph with nodes connected by two arcs can be enormous, covering a large semantic expanse, enclosing distantly related objects.

To return to the motor vehicle example, there is only one sense of vehicle. The list of immediate hyponyms is curiously specific: armored vehicle, bumper car, carrier, craft, military vehicle, rocket, skibob, sled, steamroller, tracked vehicle, troika, wheeled vehicle.

In following the hyponyms of vehicle down two steps, 130 synset nodes are visited, including such disparate items as arugula (a.k.a., "rocket, roquette, garden rocket"), skyrocket, mailman (a "carrier"), bobsled, luggage rack (a "carrier"), carpentry (a "craft"), bomber, warship, thruster, snowmobile, tank, chariot, scooter, welcome wagon and Typhoid Mary (a "carrier").

What you don't find is automobile or truck. That requires one more step – through the node for motor vehicle.

The experiments described in Chapter 8 test several lexicon variations. The basic version is a somewhat flattened version of the WordNet hierarchy, by associating a word with its *synsets* and its immediate hypernyms and hyponyms. It accepts all senses of the word since the inputs are clustered by topic, limiting the polysemy, an assumption widely use in Information Retrieval, and since many WordNet senses are viewed as too fine grained [MM01].

Other versions combine WordNet with a lexicon created automatically by examining a large corpus, as described in the next section. In this setting, more of the WordNet hierarchy will be traversed, but only words found in the automatically compiled lexicon will be accepted. The corpus-based lexicon is expected to be noisy, but the combination will reduce the noise. If the training corpus for the automatic lexicon is sufficiently large, the pairs of words that are in reality used to refer to one another should show up, and unusual and implausible pairs, like *vehicle* $\Rightarrow$ *chariot* should be nonexistent, or too rare.

In addition, other hand-built, and therefore accurate, data are merged with the lexicon by taking the union of them. I do this with information about nominalizations and other morphological transformations from two hand-built resources, the CELEX [CEL95] data, which is compiled from several dictionaries, and NOMLEX [MGM+98], which is a compilation of nominalizations, to try to extend the reach of WordNet. Examples of the variations are shown in section 5.4.

I don't intend this discussion to be a criticism of WordNet, which is invaluable, but to

emphasize that WordNet cannot be used as a black box module. I am seeking a method to make better use of WordNet, by mixing WordNet with other resources, whether hand-built or statistically mined.

A great deal of attention has been paid in the research community to the need for an automatic way of creating dictionaries, or thesauri. As part of this effort, the shortcomings of manually created dictionaries have been widely discussed and are well-known. They are all extremely expensive to create and to update, as the language changes. It is widely believed they are incomplete, but that they are accurate.

In the next section, I will deal with the issues of what is sufficiently large, and the effect of different methods of collecting co-occurrences described in Chapter 4.

5.3 Automatic Lexicons

Methods for automatically building lexicons has been underway for more than 20 years. Supervised methods have largely been applied to domain-specific tasks, in particular the Message Understanding Conferences (MUCs) and related research. In this scenario, people mark up a sufficiently large training corpus and an automated learning algorithm is applied to the examples in order to develop more general rules. The annotations identify text spans with the required semantic labels. These, along with a mixture of lexical and syntactic patterns, are then given to a learning algorithm. Success even in the constrained domains of the MUC evaluations was modest. The best systems were only able to fill half of the domain-specific templates. Of course, performance in this task is not solely dependent on the ability to establish semantic equivalence, but this ability is a key element.

In this thesis, I am trying to make the best use of existing resources, and therefore will limit the discussion in this section to issues related to the experiments described in Chapter 8. A large amount of research has been done in this area, but I will focus on the work described in the next section. It is representative of the large-scale efforts and the results are publicly available.

5.3.1 Dekang Lin

Lin [Lin98a] used a large number of syntactic relations that he found automatically with his rule-based parser MINIPAR [Lin98b], and grouped words into classes of similar meaning. He determined similarity by computing the mutual information of pairs of words in various syntactic relationships, like subject-verb, verb-object, with each other. This dictionary is publicly available. The intuition was that similar words would appear in similar syntactic patterns. The statistics were drawn from a 64-million-word corpus of news, from the Wall Street Journal, the San Jose Mercury and the AP.

He extracted "dependency triples" from the parsed corpus, which are formed from syntactic structures. Each consists of a three-tuple, a head, a relationship type and a modifier, or argument. For example, from the sentence, "Air France Concorde flights were immediately suspended", he obtains:

(suspend obj flight)
(flight n-mod Air France Concorde)

Two features are drawn from each triple, one from each word, in order to create a large feature vector for each word in the corpus. The extracted features involving the word flight would be:

flight, n = { obj-of(suspend) }
flight, n = { nmod-by(Air France Concorde) }
suspend, v = { obj(flight) }
Air France Concorde, n = { mods(flight) }

The entries in the feature vectors are counts of these features, and the similarity between two words is computed as the similarity of their vectors under mutual information.

Mutual information I is computed in the standard way. Given a syntactic relationship, r, and words w_i and w_j,

$$I(w_i, r, w_j) = \log \frac{p(w_i, r, w_j)}{p(w_i, r, *)p(*, r, w_j)},$$

where "*" indicates all w_k occurring in the corresponding slot with relationship r.

Once all the mutual information computations are done, the similarity between two words w_m and w_n is calculated over all the pairs of $I(w, r, w')$ that are positive.

	Nouns	Verbs	Adjs
WordNet	109,195	11,076	19,358
Lin	18,009	3,687	5,491

Table 5.2: Numbers of entries for WordNet and Lin's dictionary.

$$\frac{\sum_{(i,j,r)} I(w_i, r, w') + I(w_j, r, w')}{\sum_{i,r,k} I(w_i, r, w_k) + \sum_{j,r,k} I(w_j, r, w_k)}$$

For each entry, the dictionary lists the 200 words with the top similarity values. I rewrote entries into a format to allow its merger with a similarly reformatted WordNet and set a minimum similarity value of 0.10, accepting all pairs that equaled or exceeded that threshold. The merging was done by taking the intersection of the entries for a word in Lin's lexicon and the expanded hypernyms and hyponyms from WordNet. Where Lin had no corresponding entry to WordNet, the WordNet information was used alone. Table 5.2 shows that I took 27,000 entries from Lin's dictionary. These amount to 16%, 33% and 28% of the entries in the **NIA**-WordNet lexicon.

Lin adopted a number of strategies to maintain the quality of the input. He used only sentences that were at most 25 words long, and rejected any parses that were incomplete. Still, a number of problems are apparent.

First of all, the parsed input to Lin's system is only as good as the parser. MINIPAR is a dependency parser and Lin [Lin98c] reports results on five relationships, subject, complement, prepositional phrase attachment, relative clause and conjunction. The average accuracy of these five were precision of 80.5 and recall of 70.6[4] This means that the input to the system will be quite noisy.

And second, as Lin himself points out, "a threshold is required so that all similarity values lower than the threshold are considered to be 0. Since the threshold is uniformly applied to all the words, it is impossible to use it to separate good similar words from bad ones." In other words, a threshold that is too lenient in one case, is too strict in another.

[4]Precision is correct system responses over total system responses, and recall is correct system responses over total in the answer keys.

We chose a threshold of 0.10 by examination of the entries for about 20 words, and found that the entries >= 0.10 provided relatively high precision items. Erroneous additions to the *background* in our system have the effect of spreading out, contaminating the remaining comparisons.

Although Lin's dictionary has far fewer entries than WordNet, they tend to cover the most frequent words and have a disproportionate effect. In cases where there is no Lin entry, the WordNet synsets are used.

5.3.2 An Evaluation

Evaluation of unsupervised semantic discovery systems is also a difficult undertaking, and is beyond the scope of this thesis, but Curran and Moens [CM02] conducted a very interesting comparison of automatic dictionary methods and offered some interesting observations on the problem.

They compared systems based on contexts of varying complexity, from a simple window of a few words immediately to the right and left of the target word, to use of surface parsers such as SEXTANT[Gre94], CASS[Abn96] and MINIPAR[Lin98b], with MINIPAR being the most elaborate. They also implemented a simple n-window system, in which they considered n words to the left and n words to the right of each target word. With these tools, they constructed tuples like Lin's, i.e. $(word_i, relationship, word_j)$, and from these features. They used a weighted Jaccard measure for similarity:

$$\frac{\sum_{a \in atts(w_m) \cup atts(w_n)} min(wgt(w_m, a), wgt(w_n, a))}{\sum_{a \in atts(w_m) \cup atts(w_n)} max(wgt(w_m, a), wgt(w_n, a))}$$

The weighting is a t-test between the joint distribution of a word, w and its attribute, a, with the independent distributions of each:

$$wgt(w_i, a_j) = \frac{p(w_i, a_j) - p(w_i)p(a_j)}{\sqrt{p(w_i)p(a_j)}}$$

One of the purposes of their work was to compare the trade-offs among the different systems. In one test on a corpus of 150 million words, they note that MINIPAR took 74 hours, SEXTANT, 2.6 hours, and the 1-window method, just seven minutes. The results were remarkably similar. Of the top 10 similar words, Curran and Moens found that MINIPAR

had a precision of 0.405, SEXTANT, 0.39 and the 1-window method, 0.37. The gold standard they used was a combination of available hand-built resources, although the paper states that manually built thesauri are inadequate:

> "Unfortunately, thesauri are very expensive and time-consuming to produce manually, and tend to suffer from problems of bias, inconsistency and lack of coverage. In addition, thesaurus compilers cannot keep up with constantly evolving language use and cannot afford to build new thesauri for the many subdomains that information extraction and retrieval systems are being developed for."

Despite these misgivings, they combined three hand-built resources: the Macquarie Thesaurus[R.L90] and Roget's Thesaurus [Rog11] and the Moby Thesaurus[War96]. Seventy words were selected at random from WordNet. Then, a huge gold standard was created by taking the union of the synonyms to those 70 words from the three manually built resources. The total number of synonyms was 23,207. In conclusion, the authors say that the simple, but efficient algorithms, like the 1-window method, may equal or outperform existing systems for automatic thesaurus construction given larger corpora to work with. They suggest corpora of at least 1 billion words. They urge exploration of new unsupervised and semi-supervised methods. It is clear current methods, as approximated in their experiments leave much to be desired. Even with such an expansive target, the precision of 0.4 that they found is lackluster and suggests that existing automatically obtained resources have limited value. The next section details my efforts to extend the information in WordNet with addtiional information.

5.4 Combination Lexicon

In order to overcome some of the weaknesses in using WordNet, I experimented with additional resources for use in **NIA**. **NIA**-WordNet denotes the lexicon obtained from the WordNet hierarchy; it was built by taking for each word w_i all the immediate hypernyms (those words that are one step more general than w_i), all the immediate hyponyms (those words that are one step more specific than w_i) and all the synonyms. In creating this flat representation of *Referential Equivalents* (REQ), words that exceed a polysemy threshold,

break, burn, bust up, consume, cut to ribbons, demolish, demyelinate, destruct, devour, dilapidate, disassemble, eliminate, end, explode, fracture, get, harry, interdict, kick in, kill, lay waste to, level, overcome, rape, ruin, self-destruct, shipwreck, subvert, swallow, unmake, uproot, vandalize, wash out, wipe out

Figure 5.2: The entry for the verb destroy in the **NIA**-WordNet lexicon.

---

which is measured by the number of WordNet senses, are excluded. The threshold we used was set experimentally at seven. One reason for rewriting the data was just to facilitate additions such as adding information about nominalizations from two other manually created sources, NOMLEX [MGM+98] and CELEX [CEL95].

For example, Figure 5.2 shows the base entry for the verb *destroy*.

NOMLEX provides the allowed complements for about 1,000 nominalizations, and relates the nominal complements to the arguments of the corresponding verb. The most recent version was released in 2001. Much of the detail in this resource is ignored, but the idea is to capture nouns and adjectives that are tantamount to a verb.

In the same vein, the CELEX LEXICAL DATABASE, compiled for Dutch, English and German by the Dutch Centre for Lexical Information, culls information in those languages from a number of manually created dictionaries. The project ended in 2001. The database contains morphological and syntactic variations of words, including entries for derivational information.

These additions add the following entries to the lexicon record for *destroy*:

destroyer, destruction, destructive, destructiveness

Though small in number, they seemed to be sound, covering the most likely extensions. I tried taking these entries and fetching their WordNet synonyms, but the traversal stretched too thin. For example, destroyer leads to "ruiner", "undoer", "waster", "uprooter". Also, taking all the nominalizations associated with the equivalents produced much noise. Intuitively, *demolition* is a likely substitute for the verb *destroy*, but if all extended nominalizations are taken, then *consumer* and *consumption* are added.

In a more extensive alteration, I combined **NIA**-WordNet and Lin's dictionary. For words that appeared in both, the **NIA**-WordNet entries were first expanded to follow both hypernyms and hyponyms as far as possible and then take the intersection of that expansion and Lin's dictionary. For the remaining **NIA**-WordNet entries, they were kept intact.

For example, Figure 5.3 shows the Lin entry for *destroy*; it is extensive, and it is clearly noisy. Many of the words on the list are not related to the act of destroying, and a few are antonyms. It would clearly be troublesome to use. The words in bold face are those that overlap the basic WordNet list above.

The combination eliminates tenuous entries from both the Lin and WordNet sides. From the Lin side, there are many troublesome entries, like "ban", "begin", "buy", "carry", "cause", and "close". It is easy to see how co-occurrences containing these words could produce errors. From the WordNet entries, "consume", "devour", "end", "get", "harry", "overcome" and others are removed. While the semantic connection between these words and "destroy" is perfectly clear to a human, they could easily introduce confusion in an automatic system.

The combination, being an intersection of the two, was built using a slightly different procedure from the basic WordNet lexicon. Instead of stopping after collecting the immediate hypernyms and hyponyms, we continue to traverse the tree, with one restriction: the hypernyms are traversed separately from the hyponyms in order to avoid the collection of sibling synsets. Thus, only the hypernyms of hypernyms are collected, and then only hyponyms of hyponyms. The procedure avoids the noise in the Lin lexicon, and even adds several words that were not present before because they were two or more steps away. Figure 5.4 shows the combination entry.

Although the lexicons vary to a considerable degree, their effect on results in our experiments, detailed in Chapter 7, was disappointingly low. We determined that there were two reasons for this:

- There are relatively few instances of straightforward substitutions of one common noun by another. For example, in the cluster about the crash of the Concorde, the Associated Press article uses the word *crash* 9 times, but never a synonym – like *accident*. This is a fairly high amount of repetition for an article with only 332 content

abandon, affect, alter, attack, ban, batter, begin, black out, block, blow, blow up, bomb, **break**, break into, build, build in, build up, **burn**, bury, buy, capture, carry, cause, cease, change, char, close, collapse, come from, commandeer, confiscate, construct, **consume**, contain, contaminate, control, cost, cover, create, cripple, crush, cut, damage, decimate, delay, **demolish**, deplete, deploy, desert, design, detect, devastate, develop, die, disable, disintegrate, dismantle, displace, dispose of, disrupt, do, dump, **eliminate**, endanger, engulf, enter, equip, erode, establish, evacuate, exhaust, **explode**, fail, fight, fill, fire, flatten, flood, force, generate, get rid of, go into, gut, halt, hamper, harm, have, have left, hit, hurt, ignite, improve, increase, injure, inspect, intercept, invade, involve, jeopardize, keep, **kill**, knock, knock out, lead to, leave, leave behind, **level**, like, locate, loot, lose, maintain, modernize, near, need, occupy, operate, overrun, overturn, own, paralyze, penetrate, phase out, preserve, prevent, produce, protect, provide, purchase, put, raid, ransack, ravage, raze, rebuild, recover, reduce, remove, repair, replace, restore, return to, reverse, rip, rock, **ruin**, sabotage, save, scorch, scrap, seal, search, seize, sell, send, set, set up, sever, shake, shatter, ship, shoot, shoot down, shred, shut, sink, smash, steal, stop, storm, strengthen, strike, support, surround, sweep away, sweep through, take, take away, take over, target, tear, test, threaten, topple, torch, trap, trigger, try, turn over, undermine, **uproot**, use, violate, weaken,**wipe out**, withdraw, wound, wreck

Figure 5.3: The entry for the verb destroy in Lin's Dictionary. Words in bold denote an overlap with **NIA**-WordNet

charge, crush, deracinate, dismantle, displace, eradicate, even, even out, exterminate, extirpate, point, pull down, pull down, rase, raze, root out, take down, take down, tear down, tear down

Figure 5.4: The entry for the verb destroy in the combination lexicon.

---

words, of which 180 were singletons. At the same time, there were 14 occurrences of the word *Concorde*, none of which could be linked to 6 occurrences of the synonyms *airplane*, *jet* and *plane*.

- The lexicon also fails when references are not straightforward, but elliptical, or idiomatic, or even metaphorical, as is the case in a quote in the Associated article on the Concorde, from the newspaper Le Figaro, referring to the aircraft as a "beautiful white bird."

It seems more important to be able to identify common noun references to named entities than common nouns to common nouns. This became clearer as my experiments proceeded and put me in a difficult position. As I have said, there are many available tools to identify and classify named entities, including linking shorthand names to the full name – like *Columbia University* to *Columbia*. This is a key goal of the ACE program, discussed at the beginning of this chapter, but the tools are not available nor do they seem to have reached the level of performance needed here. On the other hand, my effort to build a tool to identify *Columbia* as *the university*, *the institution* or *the school* would have meant a thesis within a thesis, and therefore this remains a limitation of my work.

5.5 Word Content

A novel aspect of this thesis the use of corpus statistics and machine learning to measure how general a word may be, or from the opposite perspective, how much content a word carries independent of surrounding words. Words with a lot of content should carry more weight than more general, vague and functional words. By zeroing in on the contentful words, the system can form better co-occurrences. This effort is a step toward finding the

core vocabulary for a given topic or cluster of documents.

In Information Retrieval (IR), the importance of a word is normally measured by $TF * IDF$. The TF factor stands for term frequency, and reflects the intuition that the repetition of a word in a document or a cluster, indicates its importance, or its closeness to the *aboutness* of the discussion. The IDF factor, inverse document frequency, reduces the weight of words that are often found in many documents. It reflects the intuition that words found in many documents will not help distinguish the kind of document being searched for. There are two weaknesses in the adoption of $TF * IDF$ for our purposes.

For one, since we are dealing with news articles, there are certain topics and themes, like *murder*, that repeat, making for very high DF values, and for another, some infrequent words are themselves either very vague or very dependent on the context, like *choice*. Some are almost like *empty nouns*, which in linguistics refers to nouns like *one* – the pronomial count noun. By themselves, such words lack descriptive content.

In order to identify such low-content, promiscuous words, I examined document co-occurrences and hypothesized that words with little content would be bound to a large number of contexts, and words with high content would be bound to fewer contexts. Context is considered a document in which a target word appeared three or more times. This value was determined experimentally. A larger number cut sharply into the number of documents that would have been used. Words like *murder* and *victory* would likely be essential to consider if summarizing a document or a cluster or documents, whether or not the word was frequently repeated in the text or texts. On the other hand, words like *choice* or *matter* are only likely to be important depending on the context. To refer to some event or circumstance as *the matter* is hardly any different from using a pronoun. Its meaning entirely depends on what it refers to. To get a better look at this phenomenon, I looked at the distributions of the term and document frequencies, including the *limited* frequencies, when the term appears at or above the threshold of 3, for this study. I also looked at the distribution of word associations, as found by the binomial likelihood ratios. An overall measure, which I call promiscuity, was obtained by combining the different statistics with rule-induction algorithms that were trained on a sample of 1,000 manually classified words. Table 5.3 shows the distribution of 10 common nouns from the Associated Press portion of

word	TF	LTF	DF	LDF	promiscuity
murder	9420	3368	5683	876	0.0
weapon	18162	10496	8399	2161	0.0
injury	13351	3626	9055	931	0.2
victory	20803	6364	13429	1753	0.2
car	20346	10266	10356	2121	0.3
discussion	4912	382	4178	111	0.7
choice	5916	568	4982	161	0.8
idea	8815	976	7205	277	1.0
matter	10572	748	9024	213	1.0
thing	27354	5204	20043	1450	1.0

Table 5.3: A comparison of nouns in the Associated Press articles in the AQUAINT corpus. TF and DF are term frequency and document frequency respectively. LTF and LDF are limited term frequency and limited document frequency. The limitation here are counts of the words only in documents where the word appears three or more times. The last column shows the promiscuity score – a combination of classifiers separating words with higher content, i.e. words that appear in a small number of contexts.

---

the AQUAINT[5] Corpus. Note that zero-promiscuity words carry the most content.

Table 5.4 shows the features that I used to determine promiscuity. To capture interactions between the features, I tried several machine learning algorithms over a set of 1,000 examples that I marked by hand. The experiments will be detailed in Chapter 7, but I am using a combination of the results of two learners – these are Ripper [Coh95] and C4.5 [Qui93] . We separated the learning over the different parts of speech. For nouns, we separated the cases of *Nouns X Nouns*, *Nouns X Verbs*, and *Nouns X Adjectives*. For verbs, we took *Verbs X Verbs* and *Verbs X Nouns*, and for adjectives, *Adjectives X Adjectives* and *Adjectives X Nouns*. The final decision is based on combining the votes of the conclusions of the learners (a total of six for nouns and four for verbs) over the relevant matchups for

[5]Advanced Question Answering for Intelligence

Feature	Description
Focus	dispersion of w_i across documents
Mean LR	average likelihood ratio between w_i and w_j
Variance LR	variance of likelihood ratio w_i and w_j
Exp Val	expected value of w_j over documents with w_i
Variance of Exp Val	variance of expected values of w_j
TF	count of occurrences of w_i
focused TF	count of w_i in where frequency $> T$
DF	count of all documents with w_i
Focused DF	count of all documents with T of w_i
Assoc	count of all w_i and w_j co-occurrences
Sig Assoc	count of significant w_i, w_j co-occurrences

Table 5.4: Features used for classifying promiscuous words. w_i is a target word that occurs frequently enough to be seen T times in some number of documents. w_j is any of the words that co-occur with w_i.

the words, normalized to a score between 0..1, with 0 meaning not promiscuous.

5.5.1 Likelihood Ratios

The underlying statistic used in the promiscuity computation above is the binomial likelihood ratio [Dun93]. I use the likelihood ratios to determine the strength of the association between words, and then examine the patterns of association, as explained in the preceding section. For k successes in n trials, it is:

$$\lambda = \frac{max_p(L(p,k_1,n_1)L(p,k_2,n_2)}{max_{p_1,p_2}L(p_1,k_1,n_1)L(p_2,k_2,n_2)}),$$

where

$$L(p,k,n) = p^k(1-p)^{n-k}.$$

In the problem here, the maximum likelihood estimates are used for the values p, p_1 and p_2, so that for two words in a corpus of N words, the computation is based on:

$$p = \frac{Count(word_2)}{N},$$

$$p_1 = \frac{Count(word_1 word_2)}{Count(word_1)}, and$$

$$p_2 = \frac{Count(word_2) - Count(word_1 word_2)}{N - Count(word_1)}$$

where $Count(i)$ denotes the frequency of i.

The ratio compares the binomial distributions when the occurrences of two words are independent, in terms of conditional probabilities, $p(word_2|word_1) = p(word_2|\overline{word_1})$, and when they are dependent, $p_1(word_2|word_1) \neq p_2(word_2|\overline{word_1})$.

An important feature of likelihood ratios is that the quantity $-2log\lambda$ is asymptotically χ^2 distributed. The likelihood ratio tests do not depend on the assumption of normality as do many other statistical tests, but the χ^2 critical values can be used with the degrees of freedom set at the difference in the number of parameters, or in our case $df = 1$, which has a critical value of about 6.6. Dunning argued that this statistic is desirable because it does not make the assumption of normality and because it is more accurate with low-frequency objects.

The ordered tables of likelihood ratio values is interesting in itself. The top 15 for the set for the verb *destroy* is:

destroy, weapon, inspector, missile, fire, sanction, Iraqi, council, house, destruction, tornado, embryo, biological, damage, chemical, document, storm, resident, burn, warhead, long-range, certify, expert, mass, chemical, cooperation, inspection, blaze

Entries for verbs tend to show associated nouns. Note that nouns far outnumber verbs in counts of parts of speech. Since many nouns are rather sparse, they tend to result in higher values. On the one hand, this gives us some information about nominalizations, but

on the other hand also reflects transitory news interests when working in the news domain – as can be seen by the inclusion of words that fit the articles about the Iraqi weapons issue.

Returning to the example of *cat* and *pet* earlier in this chapter, the top 15 words associated with the noun *cat* are:

animal, lynx, pet, dog, fur, plague, veterinarian, monument, quarantine, rat, owner, ad, bite, tiger, biologist, species, cage, mouse

For the noun *pet*:

dog, animal, cat, owner, quarantine, fur, rabies, tiger, store, veterinarian, ad, deer, product, kennel

Using this data would allow for links from a cat to a pet, filling in the omission discussed at the beginning of this chapter, but it includes spurious entries – as much automatically collected does. The association between a cat and an owner is correct, but misleading for my referential purposes. These entries resemble relevance feedback, or topic signatures [LH00]used in summarization, but our intent is to use them in combination with other resources, as we stated above.

5.6 Conclusion

My efforts in semantics focus on trying to collapse individual tokens into equivalence classes of words that are likely to refer to one another (*REQ*s), and to ignore or reduce the weight on words that are too vague or general to convey content on their own, our promiscuous words.

The *REQs* aim to identify some of the references, namely cases where two different common words refer to the same underlying object or event. The use of vague words and low-content words can result in misclassification.

But these are both only first steps. Consider the verb *buy*. According to $TF * IDF$ metrics, this word would be uninformative. One of the forms of *buy* is found in 10,263 documents in the APW portion of the AQUAINT corpus. Out of a total of 146,709 unique words (uninflected forms), only 344 of them occur in 10,000 or more documents. By any

computation of $TF*IDF$ *buy* would not count in the weighting of a passage. Although the word is quite frequent, it also carries a very specific meaning. My method of promiscuous words imposes zero penalty on the verb *buy*.

But an automatic system still cannot draw any inferences that are easy for people to handle. Suppose that we have a passage containing this: "X bought dinner" in the background, and that we encounter a passage in a new document saying, "X ate dinner," or "X cooked dinner," or "X had dinner." The fact is that people effortlessly compute that buying dinner leads to eating or having dinner, sometimes implying that there is cooking dinner before eating it. My system would certainly label the second passage as "new", since it does not have any mechanism to draw inferences.

Despite the limitations, the experiments described in Chapters 7 and 8 will show small, but consistent improvements from using WordNet information, in particular by combining it with other resources providing information about morphological variations, in particular nominalizations.

Chapter 6

Context

Up to now the discussion has focused on the need for a fine-grained analysis of the input texts, which I've called the *Micro View*. The word pairings called co-occurrences which are formed within clause boundaries are a key facet of the *Micro View*.

It was apparent from early in the experiments that whole segments of articles – sometimes many sentences long – were often novel. These are often centered around some new entity and easily identifiable by the first appearance of proper noun phrase – a name of a person, place or organization.

Figure 6.1 shows a typical pattern of novelty in an article on America Online's legal fight against spammers. The article is one of the pairs in the development corpus for this work (see Chapter 3). The shaded words show what the annotators agreed was new information. They include four subsentential chunks, including the facts that the current case was one of a series of lawsuits and that AOL never lost against spammers. They also include two whole sentences covering where the current suit was filed and details on how spammers operate. Finally there is a four paragraph chunk of text that adds a lot of new detail about Michael Levesque, an accused spammer.

In this chapter, I briefly turn away from the *Micro View* to explore how the system can capitalize on the observation that new information is packaged in chunks of sentences and larger. The Text Retrieval Conference's Novelty Track offered a convenient way to test against such long spans, since its gold standard judgments are made on sentences and sentences only. The sentence targets made a poor target for the *Micro View* system, but

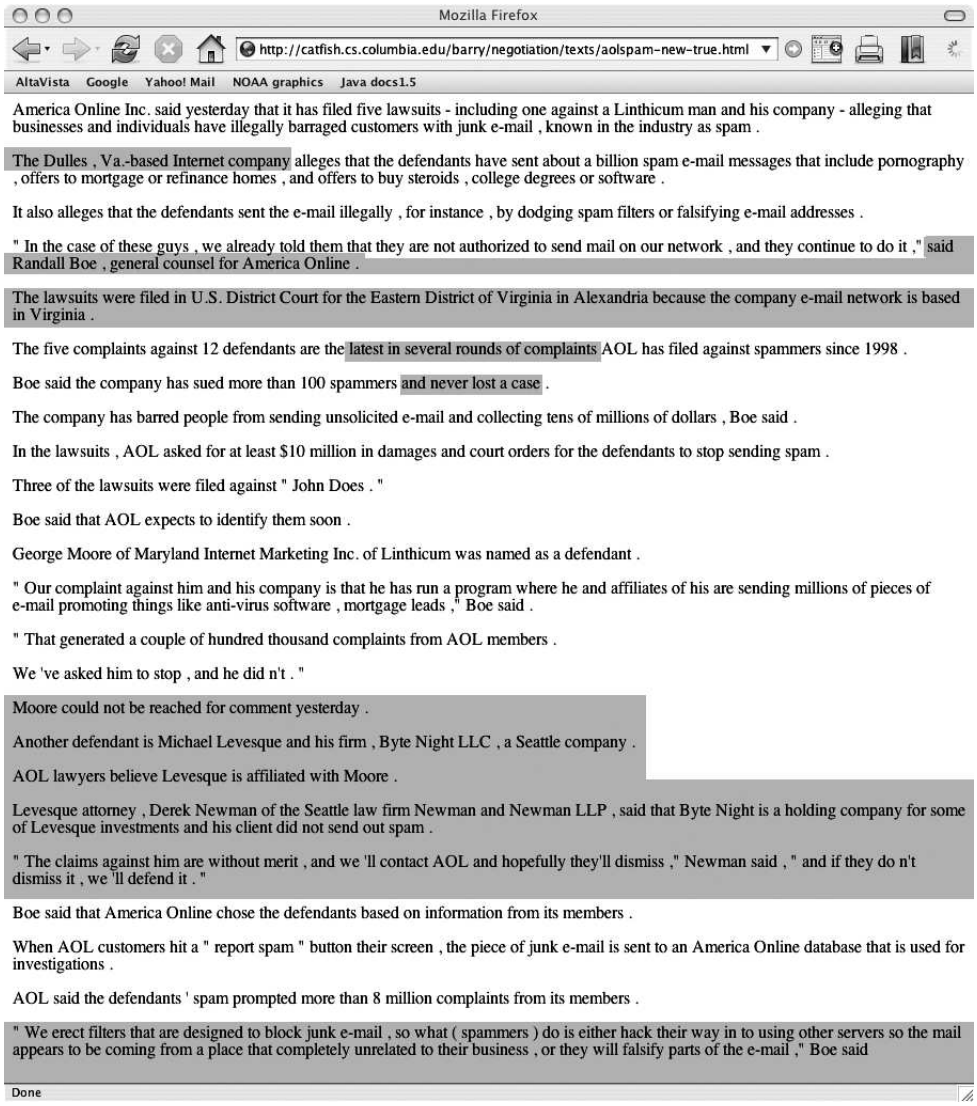

Figure 6.1: This article on America Online's fight against spam shows a typical pattern of new information in different sized pieces – from short phrases to segments several paragraphs long. The shaded parts are the areas that two human annotators agreed was novel when this article was compared to another on the same event.

they were an opportunity to test some ideas for a *Macro View*. I wrote a program, called SumSeg, to operate specifically on the TREC data, and it achieved the highest precision scores in the Novelty Track evaluation. Several characteristics derived from SumSeg were then incorporated in **NIA**.

In order to determine whether information within a sentence has been seen in documents read so far, SumSeg integrates information about sentence context, novel words within the sentence, and named entities all as part of a specialized learning algorithm. A key property of the algorithm is the ability to track focus to determine sentence context, identifying and tracking segments of novel information. In addition, the program also tested the effect of trying to eliminate low-content words, or promiscuous words, which were discussed in Chapter 5.

The Novelty Tracks in 2003 and in 2004 were divided into four tasks; Task 1 and Task 3 incorporate retrieval, requiring submissions to locate the relevant sentences before filtering them for novelty. Tasks 2 and 4 are new-information detection alone, using the relevant sentences selected by humans as input. Since my interest is in novelty detection, I chose to concentrate on Task 2[1].

My Novelty Track submission was designed to test specialized learning mechanism, which learned settings to target either high precision or high recall. The five submitted runs were:

Prec1 A run to obtain moderately high precision, retaining reasonable recall.

Prec2 A run to obtain high precision, with little attention to recall.

Recall A run which weighted precision and recall equally.

Cosine A baseline run of a standard vector-space model with a cosine similarity metric.

Combo A composite submission using the intersection of the Recall and Cosine Runs.

[1]Since Tasks 1 and 3 depended on relevance decisions, I did not do them; Task 4 was similar to Task 2, in that both have the human annotations as input. For Task 2, that's all participant get, but in Task 4, they also receive the novel sentences from the first five documents as input. I felt that I would learn as much from the Task 2 as from both Task 2 and 4.

In addition to developing a mechanism for tracking focus for novelty detection, I also experimented with reducing the weights of *promiscuous* words. While many researchers use inverse document frequency (IDF) to identify words that appear frequently throughout a corpus, I also noted that some words can appear to be distinctive of a particular document and yet, still convey little content. Such words (e.g., "keep," "take," "type") typically co-occur with a much wider variety of words than do most.

The submissions to TREC all included the *promiscuous* words feature, but later experiments found that their contribution to precision was not substantial. Future extensions that could increase their impact are discussed later in this chapter.

The next section will discuss the Novelty Track; Section 6.2 will detail SumSeg, and Section 6.3 will review the experiments and their outcomes.

6.1 Novelty Track

Much of the work in new-information detection has been done for the Novelty Track, which were conducted as part of the Text Retrieval Conference in 2002, 2003 and 2004. The task is related to first story detection, which is defined on whole documents rather than on passages within documents. In Task 2 of the Novelty Track, the inputs are the set of relevant sentences, so that the program does not see the entire documents.

6.1.1 Precision

At all three Novelty Track evaluations, it is clear that high precision is much harder to obtain than high recall. Trivial baselines – such as accept all sentences as novel – have proven to be difficult to beat by very much. This one-line algorithm automatically obtains 100% recall and precision equal to the proportion of novel sentences in the input. In 2003, when 66% of the relevant sentences were novel, the mean precision score was 0.635[2] and the median was 0.7. In 2004, 41% of the relevant sentences were novel, and the average

[2]One group appeared to have submitted a large number of irrelevant sentences in its submission, since it obtained relatively high recall scores, but very low precision scores, causing the combined score to drop below 0.66. The average precision of all other groups is about 0.7.

precision dropped to 0.46. The mean precision was also 0.46. Meanwhile, average recall scores across all submissions actually rose to 0.861 in 2004, compared with 0.795 in 2003. In terms of a real world system, this means that as the number of target sentences shrank, the number of sentences in the average program output rose.

6.1.2 Sentences

The Novelty Track is organized on the basis of sentence units, and an exhaustive comparison of a each sentence against the others was the standard approach. However, two sentences may be partly similar and partly different. So a system could classify one as novel because part of it is new, or as old because part of it is not new. Averaging the two components would seem to invite some randomness in the results, and this is what seems to have happened at the Novelty Track.

Normally, in Information Retrieval tasks, stricter thresholds result in higher precision, and looser thresholds, higher recall. In that way, a system can target its results to a user's needs. But in new-information detection, this rule of thumb fails at some point as thresholds become stricter. Recall does fall, but precision does not rise. In other words, there seems to be a ceiling on how far precision can be raised. We will see that behavior in Section 6.3 when we test my baseline vector-space system.

In earlier years, several of the participants noted that their simpler strategies produced the best results. For example, in 2003, the Chinese Academy of Sciences [SpZ+03], noted that word overlap was surprisingly strong as a similarity measure. Also in 2003, the University of Iowa [ESL+03] achieved the highest precision scores just by counting previously unseen words. As we have seen above, this strategy was incorporated in several other systems for 2004, including mine. This strategy compares words in a sentence against all previous seen words used and thus, avoids direct comparison of one sentence against another.

In Chapter 4, I discussed why a sentence-by-sentence comparison is clearly not the optimal operation for establishing novelty. This situation motivated the construction of my Novelty Track experiments.

6.2 System

SumSeg was built with the Novelty Track in mind. The goal was to look at ways to consider longer spans of text than a sentence, while avoiding sentence by sentence comparisons.

In the Novelty track, the relevant sentences are presented in natural order, i.e. by the date of the document they came from, and then by their location in the document.

SumSeg:

- Calculates a weighted sum of novel terms, which are terms that have not been seen in any previously scanned article. The weights are learned automatically.
- Maintains a focus variable, which indicates whether the previous sentence is novel or old. Thresholds determine whether to continue or shift the focus. These are also learned automatically.

Figure 6.2 shows the system architecture. All input documents are fed in parallel into a named-entity recognizer, which marks persons, organizations, locations, part-of-speech tags for common nouns, and into a finite-state parser, which is used only to identify sentences beginning with subject pronouns. (There was no need for the more powerful parsing described in Chapter 4 for this task.) The output from the two preprocessing modules are then merged and forwarded to the classifier.

The classifier reads a configuration file that contains a set of weights to apply to novel words, assigning different weights to different parts-of-speech (e.g. person, common noun, verb) and then scans each sentence. The weights used in the 2004 Novelty Track were learned over the Novelty Track data for 2003.

For each sentence, the system computes the amount of novelty from the weighted terms in a sentence and compares that to a learned threshold; it classifies the sentence as novel if it exceeds the threshold. It also stores the classification, i.e. novel or not, in a focus variable. If the novelty threshold is not met, the system checks the series of thresholds described below, and possibly classifies some sentences with few content words as novel, depending on the status of the focus variable. SumSeg tries to cover all cases of changes in focus, and to test these in the order that allows the system to make the decision it can be most confident

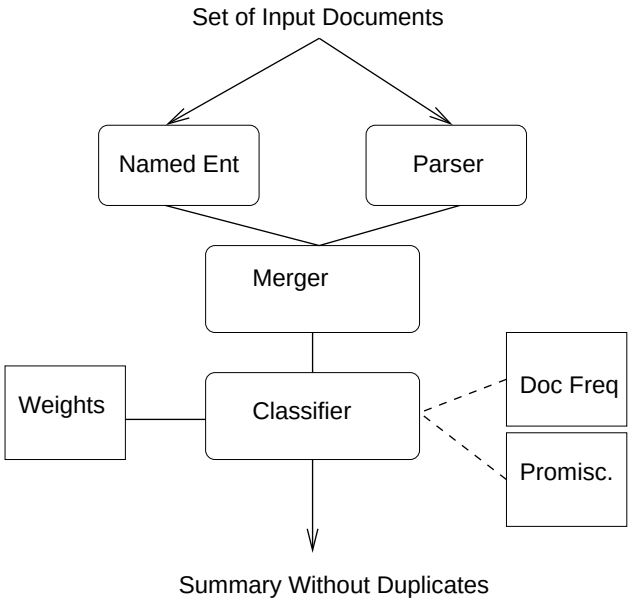

Figure 6.2: Architecture of SumSeg for the Novelty Track.

about first. Thus, when it finds a named-entity new to the discussion, it is fairly confident that it has found a sentence with new information. It can classify that sentence as new without regard to what came before. But, when it finds a sentence devoid of high-content words, like "She said the idea sounded good," it will follow the classification of the previous sentence. If the antecedents to *she* or *idea* are novel, then this sentence must also be new. The series of learned thresholds are tested in a cascade to maximize the number of correct decisions over the training cases, in hopes the values will generalize to cover unseen cases.

The SumSeg approach is unique in representing and maintaining the focus. The idea stems from the fact that novelty often comes in bursts, which is not surprising since the articles are composed of some number of smaller, coherent segments. Each segment is started by some kind of introductory passage, and that is where the *novel* words are likely to be. Consequently, the presence of novel words are the primary evidence that the entire segment is likely to contain more novel material. Subsequent passages are likely to continue the novel discussion whether or not they contain novel words. They may contain pronomial references or other anaphoric references to the novel entity.

Thus, the classifier puts each sentence through the following tests, using the learned

thresholds and weights described below.

1. It checks if the sum of the weights of the novel content words (including named entities) exceeds a threshold, T_{novel}. If it does, the sentence is considered novel. If the previous focus was old, this indicates the focus has shifted to a novel segment.

2. If the sum of the weights of novel words does not exceed T_{novel}, it compares the weight of the already-seen content words against a separate threshold, T_{old}. If the threshold is passed, the sentence is considered old. If the previous focus was novel, this means the focus has shifted to an old segment.

3. The next test is twofold:

   (a) If the sum of the weights of old content words and novel content words is below a threshold, T_{keep}, the prior focus, novel or old, is kept because such a passage is not likely to indicate a segment shift.

   (b) If the first noun phrase that is not contained in a prepositional phrase is a third person personal pronoun, I assume the prior focus, novel or old is kept. Pronouns are known to signal that the same focus continues [GS86].

4. The default is to continue the focus, whether novel or old.

I examined the 2003 Novelty Track data and found that more than half the novel sentences appear in sequences of consecutive sentences (See Table 6.1). This creates an opportunity to make principled classifications on some sentences that have few, if any, clearly novel words, but continue a new segment. The use of a focus variable handles these cases.

6.2.1 Learning Weights and Thresholds

In all, the system uses 11 real-valued weights and thresholds. The learning mechanism was designed to learn optimal values for these, and in particular, to target either high recall or high precision. As noted above, precision was much more difficult than recall. In an overall task like summarization, precision is much more important.

Length of Run	Count
1	1338
2	421
3	132
4	72

Table 6.1: Novelty often comes in bursts. This table shows that 1,338 of the novel sentences in the 2003 evaluation were singletons, and not a part of a run of novel sentences. Meanwhile, 1,526 of the sentences were part of runs of 2, 3 or 4 sentences.

In SumSeg, the classifications of the examples is made on line; the feature values for training instance i depend in part on the decision made on those for training instance $i-1$. Popular supervised learning methods like naïve bayes, or k-nearest neighbors, or decision trees use static feature vectors, and thus are inappropriate for this task. I used a randomized hill-climbing approach to find effective parameters for the system – borrowing from neural nets and genetic algorithms (See Figure 6.3) – and avoid local minima. The evaluation, or fitness function, is the Novelty Track score itself, and the training data was the 2003 Novelty Track data.

Changes to the hypothesis are selected at random and evaluated. If the change does not hurt results, it is accepted. Otherwise the program backtracks and chooses another weight to update. At first, I required the new configuration to produce a score greater than the previous one before I accepted it. But I altered this to accept configurations that produce scores equal to the previous one. The choice of which weight to update is made at random, in an effort to avoid local minima in the search space, but with an important restriction: the previous n choices are kept in a history list, which is checked to avoid re-use. This list is updated at each iteration.

The configurations usually converge well within 100 iterations. I experimented with ways to initialize the starting values. I first tried handpicked values and then uniform weights, but found convergence was usually faster with random starting values.

In training on the 2003 data, the biggest problem was to find a way to deal with the large percentage of novel sentences. About 65% of the instances are positive, so that a

1. $T \leftarrow$ Training cases
2. Initialize weights, W, history, H
   - $W \leftarrow$ random values
   - $H \leftarrow \emptyset$
   - $Best \leftarrow$ apply W to T
3. Change random weight $w \in W$
4. $Current \leftarrow$ apply W to training cases
   - If $Current \geq Best$
     - $Best \leftarrow Current$
   - Otherwise
     - Undo 2
5. Update H, Go to step 3

Figure 6.3: The learning algorithm uses a randomized hill climbing approach with backtracking

random system achieves a relatively high F-measure by increasing the number of sentences it calls novel – until recall reaches 1.0. Another strategy would be to exclusively choose the sentences in the first document, achieving a high precision – more than 90% of the relevant sentences in the first document for each topic were considered novel.

In the Novelty Track, the F-measure was set to give equal weight to precision and recall, but I wanted to be able to coax the learner to give greater weight to either precision or by adjusting the F-measure computation:

$$F = \frac{1}{\frac{\beta}{prec} + \frac{(1-\beta)}{recall}}$$

β is a number between 0 and 1. The closer it gets to 1, the more the formula favors precision.

I chose whether to emphasize precision or recall by altering the value of β. For my *Prec1* run, I set β at 0.7; for *Prec2*, it was set at 0.9, and for *Recall*, it was set at 0.5.

The design was motivated by the need to explore the problem more fully and inform the algorithm for deciding novelty as much as to find optimal parameters for the values. Thus, I wanted to be able to record all the steps the learner made through the search space, and

to save the intermediate states. At times, the learner would settle into a configuration that produced a trivial solution, and I could choose one of the intermediate configurations that produced a more reasonable score.

6.2.2 Word Content

SumSeg used two strategies to identify the more important content words in the input sets. The first was a variation of the $tf * idf$ metric and the second was the promiscuity metric described in Chapter 5. I felt that the two measures would be complementary and enhance performance in different ways. Both metrics are used to scale the value of content words when summing the total novelty, or familiarity, of a sentence.

In order to emphasize words that are important to the input document set, SumSeg used a metric that is similar to $tf * idf$, but computes the tf factor over all the documents in an input set. Document frequencies are taken from the underlying AQUAINT corpus, from which the input sets in the Novelty Track evaluation were drawn. The system counts the lemmas, the uninflected word roots, to combine the obvious morphological variations and used a log scale for the document frequencies to flatten the values. The metric is used as a weight, W, that based on the product of the two values:

$$W = (1 - (\frac{1}{log(df_{set})}))(\frac{1}{(log(df_{corpus}))})$$

For words in the underlying corpus, the value W becomes the weight of the word. For other words, mainly named-entities, the value is 1.

I am also interested in finding a way to avoid false novelty readings because of chance occurrences of words that are vague. In most Information Retrieval contexts, stop lists or $tf * idf$ weighting is used. It is the *idf* factor, the inverse document frequency, that balances the frequency within a document. For example, the verb *say*, or one of its inflected forms, occurs several times in almost every news article, so it is not given much weight, because it has a high *idf*: it appears in 181,142 of 200,985 articles in the New York Times portion of the AQUAINT corpus, or in more than 90% of the articles.

However, there are other words that I would like to ignore because they are either too vague, or more functional that content-bearing. For example, the word *idea* could be

Nouns		Verbs		Adjs	
way	1.0	use	0.75	important	1.0
use	1.0	turn	0.75	human	1.0
type	1.0	try	0.75	high	1.0
time	1.0	stay	0.75	hard	1.0
thing	1.0	show	0.75	great	1.0

Table 6.2: A sampling of promiscuous nouns, verbs and adjectives that were found to be used in too many contexts to convey much meaning on their own.

inserted or removed from many contexts without changing the aboutness of the passage, yet it is a relatively rare word, appearing in only 30,349 documents in the same corpus.

The *promiscuous* words metric finds words that are strongly associated with many different contexts. The goal is the same as my use of document frequencies, but the method for identifying these words is their contextual distribution.

SumSeg used a threshold of 0.55, which was chosen experimentally. Here, if the value exceeds that threshold, the word is eliminated from consideration. The noun *idea* gets a promiscuity score of 1.0, and thus would be ignored when the system makes its classification. Table 6.2 shows a sample of such words.

I do not argue that these words have no content or meaning, but that they are either intrinsically vague or are commonly used in structures that are semantically dominated by another word, like "a type of vehicle". The word *type* provides information about the object, but *vehicle* is the word I want.

6.2.3 Vector-Space Module

In addition to SumSeg, I also implemented a vector-space approach as a baseline – the *Cosine* run. I tested the vector-space system alone to contrast it with the SumSeg system, but I also tested a version which integrated both.

The vector-space module assigns all non-stop-words a value of 1, and uses the cosine distance metric to compute similarity.

$$Cos(\vec{u}, \vec{v}) = \frac{\sum_i u_i v_i}{\sum_i u_i^2 \sum_i v_i^2}$$

and

$$Novel(s_i) \begin{cases} true & if Cos(s_i, s_j) < T, for j = 1 \ldots i-1 \\ false & otherwise \end{cases}$$

As each sentence is scanned, its similarity is computed with all previous sentences and the maximum similarity is compared to a threshold T. If that maximum exceeds T, it is considered novel. I chose the value of T after trials on the 2003 Novelty Track data. It was set at 0.385, resulting in a balanced system that matched the results of one of the strongest performers at the TREC evaluations that year.

On the 2003 data, when I set T at .9, I found that I had a precision of .71 and a recall of 0.98, indicating that about 6% of the sentences were quite similar to some already-scanned sentence, as noted in Chapter 4. After that, each point of precision was very costly in terms of recall. My experience was mirrored by the participants at TREC 2003 and again at TREC 2004.

I considered this vector-space model to be a baseline. I also tried it in combination with the *Recall* run explained above. Because both the *Recall* and *Cosine* runs produced a relatively large output and because they used different methods, I thought the intersection would result in higher precision, though with some loss of recall.

In practice, the range of recall was much greater than precision. Judging from the experiences of the participants at TREC and my own exploratory experiments, it was difficult to push precision above 0.80 with the TREC 2003 data, and above 0.50 with the TREC 2004 data. The next section presents the results in the 2004 Novelty Track.

6.3 Experiments

6.3.1 Results from TREC 2004

The results were encouraging, and prompted the inclusion of the *Macro View* into **NIA**. It was especially encouraging that the configurations that were oriented toward higher pre-

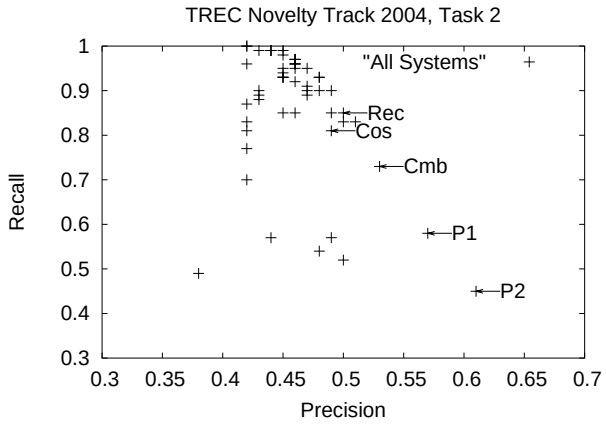

Figure 6.4: The plot shows all 54 submissions in Task 2 for the Novelty Track, with my five submissions labeled. My precision-oriented runs were well ahead of all others in precision, while my recall-oriented run was in a large group that reached about 0.5 precision with relatively high recall.

cision, indeed, achieved the best precision scores in the evaluation, with my best precision run about 20% higher in precision than the best of all the runs by other groups (see Figure 6.4). Meanwhile, the recall-oriented run was one of eight runs that were in a virtual tie for achieving the top F-measure. These eight runs were within 0.01 of one another in the measure.

Table 6.3 shows the numbers of the performance of my five submissions. *Prec1* had an F-score close to the average of 0.577 for all systems, while *Prec2* was 50% ahead of random selection in accuracy. Both my *Combo* system and my baseline *Cosine* were above average in F-measure. The emphasis on precision is justified in a number of ways, although the official yardstick was the F-measure.

First, in the larger context, **NIA** is a summarization system, where compression of the report is valuable. Table 6.3 shows the lengths of my returns. It is impossible to compare these precisely with other systems, because the averages given by NIST are averages of the scores for each of the 50 sets, and I do not have the breakdown of the numbers by set for any submissions but my own. However, size of the others' output can be estimated by considering average precision and recall as if they were computed over the total number of

Run-Id	Precision	Recall	F-meas	Output length
Prec1	0.57	0.58	0.562	3276
Prec2	0.61	0.45	0.506	2372
Recall	0.51	0.82	0.611	5603
Cosine	0.49	0.81	0.599	5537
Combo	0.53	0.73	0.598	4578
Choose All	0.41	1.000	0.581	8343
Average All Runs	0.46	0.86	0.577	6500

Table 6.3: Comparison of results of my five runs, compared to a random selection of sentences, and the overall average F-scores by all 55 submissions.

sentences in all 50 sets. This computation shows an average output for all participants of about 6,500 sentences and a median of 6,981 – out of a total of 8,343 sentences. My outputs are shown in Table 6.3. However, this total includes some amount of header material, not only the headline, but the document ID and other identifiers, the date and some shorthand messages from the wire services to its clients. In addition, a number of the sets had near perfect duplicate articles. I contend there is little value in a system that does no more than weed out very few sentences, even though they might have achieved high F-measures.

Second, my experience, and the results of other groups, shows that it is much more difficult to achieve high precision than high recall. In all three years of the Novelty Track, precision scores tended to hover in a narrow band just above what one would get by mechanically selecting *novel* for all sentences.

Finally, the F-measure is problematic in this task, as NIST concedes in its overview [Sob04], because the same score can be achieved by vastly different systems. In cases with relatively few true novel sentences, accuracy and coverage are better matched in difficulty.

6.3.2 Test of promiscuity

In addition to my official submission to TREC, I conducted a series of experiments to test the efficacy of my word-content measures – both promiscuous words and document frequencies, which are detailed in Section 6.2.2. The idea of measuring content or importance with $tf*idf$

Run ID	Precision	Recall	F-measure
Promiscuous Words Alone			
Prec1	0.57	0.58	0.563
Prec2	0.61	0.45	0.505
Recall	0.51	0.82	0.613
Document Frequency Alone			
Prec1	0.56	0.64	0.583
Prec2	0.59	0.52	0.537
Recall	0.50	0.85	0.617
No Content Measure			
Prec1	0.56	0.65	0.585
Prec2	0.58	0.54	0.544
Recall	0.50	0.86	0.619

Table 6.4: Comparison of methods of measuring low content words.

is well established in the Information Retrieval community. It was my goal to sharpen the idea by eliminating low-content words and thereby reduce false alarms in my task. Table 6.4 shows that there is more work to do in this area. When used alone, promiscuous words produce average scores very close to the scores on my submitted runs, in which I used both the document frequency metric and the promiscuous words. When compared with runs without any measure of word content, promiscuous words produces small improvements in precision at a considerably larger drop in recall. In future work, low-content words must be examined as potential anaphora. Words such as “idea” or “decision” are likely to be anaphoric, as they refer to something that previously appeared in the discourse, and ignoring them would tend to lead to lower recall. Taken at face value, they would behave like wildcards, leading to false positives and lower precision.

6.4 Conclusion

The TREC results confirm that the use of a focus variable to consider context in deciding novelty is a good approach. SumSeg achieved the top precision scores, and the program settings to do this were automatically obtained. But precision remains costly in terms of loss of recall. The efforts to weed out low-content words added slightly to precision scores, at a steep cost in recall.

In the larger realm, the results at TREC were gratifying. They confirmed that comparing two sentences against each other are not the approach in this task, and they also suggested a significant addition to **NIA** in the focus variable. The idea was incorporated into **NIA** by the addition of features to reflect the distance between the current clause and the last clause with novel elements. In addition, other features were added to reflect straightforward counts of new elements in each clause. In Chapter 7 the experiments with rule induction algorithms will show that these features contribute much to overall results.

Chapter 7

Learning

This chapter discusses the application of machine learning techniques to new-information detection. Chapters 4, 5 and 6 covered the development of various features to characterize passages as either new or old. Chapter 4 showed how **NIA** uses syntax to analyze the structure of the input documents. Chapter 5 developed methods of applying surface semantics to equate different tokens that referred to the same underlying entities. The combined syntactic and semantic processing allows for the identification of the co-occurrences needed to identify new information. These two types of analysis form the *Micro View* of the problem.

In order to have a starting point for the experiments that followed, a baseline **NIA** was constructed, using a bare-bones approach: any clause that contained a single novel co-occurrence of content words was classified as new. In addition, Chapter 6 described the SumSeg system, a sentence-based approach designed for the TREC Novelty Track that used a sum of novel content words, adjusted for informativeness, and a focus variable to include a sense of context in the novelty determination. SumSeg is the basis for the *Macro View*.

Machine learning is used as the vehicle for combining the Macro and Micro approaches, and for raising precision to more useful levels. Table 7.1 shows that both clause-based baseline **NIA** and sentence-based vector-space with a cosine metric (Cosine) approach were hardly able to be more discriminating than a take-all-passages approach. Gains in precision are very costly in terms of recall. A 7% gain in precision by SumSeg over Cosine cost a 33% loss in recall. Likewise a 4% gain in precision by Cosine over NIA cost a 7% loss in recall.

System	Precision	Recall	F-measure
Accept All	0.617	1.00	0.763
Baseline **NIA**	0.624	0.795	0.699
Cosine	0.647	0.738	0.688
SumSeg	0.692	0.492	0.574

Table 7.1: Summary of baseline results on our 31 pairs of training articles, including the system of accepting all inputs, the initial version of **NIA**, without learning, the cosine system with a threshold of .50, the SumSeg system trained on the TREC Novelty Track data for 2003. In precision, **NIA** was not significantly better than Accept All, and Cosine was only marginally better. NIA and Cosine were also indistinguishable from one another, according to the binomial test. In precision, SumSeg was significantly ahead of Cosine.

These experiments were performed on my data, collected as described in Chapter 3. While SumSeg performed well at TREC, it was not as strong on my finer-grained annotation.

Over all, this initial experiment mirrors the typical experience by participants at the TREC Novelty Track in 2003 and 2004, as seen in Chapter 1. None of the systems is able to beat the trivial accept-all approach, which is by definition a useless system that does nothing but return the input to the user.

Machine learning offers two avenues for improvement.

First, it allows for a richer representation of the input texts and then discovers the features that are the most important. Second, learning allows the *Micro View* clause-based and *Macro View* sentence-based systems to be integrated by adding features from both perspectives. In addition, each approach can be considered as a relatively weak classifier, and then combined through bagging and boosting techniques. These can often produce greatly improved results by building a number of independent classifiers and using them as a kind of committee.

The next section addresses the need for experimentation in choosing a learning algorithm for a particular kind of data. Section 7.0.1 presents the features that are extracted. I tried to be inclusive in choosing features, but to concentrate on those that would be reliably

obtained.

7.0.1 Features

NIA uses 24 features in all, combining the *Micro* and *Macro Views* of the inputs, and these are far from exhaustive. Figure 7.1 shows the complete list, including three additional features that merely identify the clause, sentence and article and that are therefore ignored when the hypothesis is built. I sought to strike a balance between reliability and expressiveness. The most complicated features, the co-occurrences, do rely on surface structure, but no effort is made to interpret the meaning or assign semantic roles. Such deep knowledge would certainly be desirable, but only if the tools to do the interpretation were sufficiently reliable.

Figure 7.1 is a list of the features extracted by **NIA**. Features 1 through 8, plus 12, 13 and 15 are from the *Macro View*, inspired by SumSeg. Features 17 through 24 are drawn from the *Micro View*, and are an elaboration of the original notions of co-occurrences in the baseline **NIA**, shown in Table 7.1. Some additional structural features were inspired by the DEMS multi-document summarizer. These are 9, 10 and 27. Feature 27 was based on the notion that different rules might be more appropriate for documents that are narrowly about a single topic than for documents that cover several events. Feature 25 is based on the sentence-based cosine-similarity metric. Feature 26 was included to see if perhaps there was a rhetorical cue for novelty.

Learning algorithms are designed to navigate in the feature space and build a hypothesis out of the most effective features. This chapter will show that noise is a problem for **NIA**, but that options offered by the learners help to create a more general, less brittle solution. These include techniques like discretizing the feature values or pruning the rules after training. Clearly, **NIA** is operating in a vast feature space. Even if all the features had only binary values, the feature space would still be large enough to find a unique mapping covering more than 16 million examples. Of course, this is a hypothetical situation, but in practice overfitting is a serious danger. In **NIA**, only three of the features are symbolic, *Quad, Subjtype and Conjunction*. The rest are numeric, many real-valued. There is also a scarcity problem as well. There are only 2,023 clauses in the 31 pairs of articles – an

average of 32.6 clauses per article. This is not an entirely comfortable number for the size of the hypothesis space.

Figure 7.2 provides an illustration of this situation. It shows a small part of the hypothesis found by the *One-R* learner in the Weka suite. *One-R*, which is often used as a simple baseline, chooses one feature and uses it to devise a rule. The feature chosen was *Cover*, the highest sentence similarity value of the current sentence with any previously seen sentence – essentially the Cosine system. Using our entire training corpus, *One-R* came up with a hypothesis composed of 141 rules all of the form below. In effect, the function representing the hypothesis divided up the domain of possible feature values into 141 narrow intervals and classified each instance according to which interval it fell. The intervals, from $0 \ldots 1$, thus average 0.00709 each – an extremely fine-grained view of a coarse measure.

It's hard to believe that such a rule of alternating narrow intervals would generalize – where 0.280 is old, and 0.284 through 0.292 is new and 0.293 is old. However, in terms of precision and recall, this system did very well, with 0.789 and 0.935, respectively, for an F-measure of 0.856 in 10-fold cross validation. This solution is drastically overfit to the training data.

In the next section, I will discuss the difficulty of choosing the best learning method, and to explain the need for an experimental strategy – to try an array of approaches, and then to find optimal parameters for the most promising approach. Following that, I will give an overview of the various learning methods we tried, including naïve bayes, support vector machines, k-star, decision trees and rule induction, and their performance on my data.

7.1 Learning Methods

The choice of learning algorithm is no trivial matter. Although there is a growing body of work in computational learning theory, there is no clear guide about what method will work best in a certain situation. In fact there are results, known as the *No Free Lunch* theorems [WM97, WM95] that state the success of learning depends in large part on our knowledge of the problem domain. According to the theorems, any two search, or opti-

1.Novwrd A count of novel content words
2.Novnns A count of novel nouns
3.Novvrb A count of novel verbs
4.Novadjs A count of novel adjectives
5.Novpeop A count of novel people
6.Novorg A count of novel organizations
7.Novloc A count of novel locations
8.Novnam A count of novel unclassified names
9.Rawsize Number of total words in the unit
10.Quad Which quadrant of the article the current unit is in
11.Clausid* Identifier
12.Subjtype Type of subject, i.e. the tag of the subject noun
13.Novdist Distance of current to closest previous unit w/novel words
14.Docname* Identifier
15.Pnov Distance between this and the closest previous novel sentence
16.Sentkey* Identifier
17.Oldbigrams Count of Co-Occurrence that were already seen
18.Newbigrams Count of Co-Occurrences that are new
19.Bestbg Highest globally valued new co-occurrence
20.Avergb Global average of new co-occurrences
21.Glodist Distance between elements of best global co-occurrences
22.Bestset Best set value of new co-occurrences, based on frequency
23.Setmean Average set value of new co-occurrences
24.Setdist Distance between elements of best set co-occurrence
25.Cover An overlap of the current sentence by any of the previous
26.Conjunction If there is a subordinate clause, what is the conj
27.Covereddoc Overlap measure of best doc similarity

(* Identifiers are, of course, ignored during the learning procedure.)

Figure 7.1: The features used in **NIA**.

$$Novel(Unit_i) = \text{True if} \begin{cases} 0 \leq U < 0.16976708946129532 \\ 0.1775632660837524 \leq U_i < 0.2669236007979576 \\ \mathbf{0.2844835146284316 \leq U_i < 0.29226932207084516} \\ \ldots \\ 0.9706651969125835 \leq U_i \leq 1.00 \end{cases}$$

Figure 7.2: Examples of the narrow slices values. This fragment of the full table of 141 rules, with each rule setting out a narrow slice of the interval of $0 \ldots 1$. The rule in bold, on the third line, shows an interval of $, 0.01$. Sentences within that interval are new, while those in the narrow bands above and below are classified as old. This is clearly an artifact of data.

---

mization, algorithms are equally likely to outperform the other on any particular problem [WM95].

> "For any pair of search algorithms, there are as many problems for which the first algorithm outperforms the second as for which the reverse is true. One consequence of this is that if we don't put any domain knowledge into our algorithm, it is as likely to perform worse than random search, as it is likely to perform better. This is true for all algorithms ..."

Different learning algorithms produce different results – even within one algorithm, there are usually a variety of options whose settings can produce substantially different results.

This necessitates experimentation, which I conducted with a comprehensive suite of learning tools, WEKA [WF00], as detailed below. In all of the tests, I used 10-fold cross validation over the annotated clauses described in Chapter 3.

7.1.1 Naïve Bayes

The naïve bayes algorithm assumes that the features are independent of one another, an assumption that many have pointed out has "no basis in reality." Despite this conceptual shortcoming, the naïve Bayes classifier is often an effective tool that compares well to more

sophisticated learning techniques on many problems, including some in natural language processing, like text categorization.

Formally, we are given a training set of pairs each consisting of a feature vector, $\vec{x}$, of length n, and a label, $c \in C$ that describe an instance of the problem and its classification. From the distribution of these feature values over the training set, we seek to be able to label unseen examples by finding the most likely classification.

$$c = \arg\max_{c_i \in C} p(c_i | x_{i,1}, x_{i,2} \ldots x_{i,n})$$

which can be rewritten, according to Bayes theorem to:

$$c = \arg\max_{c_i \in C} \frac{p(x_{i,1}, x_{i,2} \ldots x_{i,n} | c_i) p(c_i)}{p(x_{i,1}, x_{i,2} \ldots x_{i,n})}.$$

The naïve Bayes classifier makes the simplifying assumption that the elements of the feature vectors are independent of one another, allowing us to substitute the product of the probabilities for the conditional probabilities. We also drop the denominator since it does not affect the result:

$$c_i = \arg\max_{c_i \in C} p(c_i) \prod_j p(x_{i,j} | c_i)$$

The standard naïve Bayes classifier, like the default configuration in Weka, uses a Gaussian probability distribution to compute $p(x_{i,j}|c_i)$, where x is a continuous value (as are most of the features developed by **NIA**). For discrete values, the classifier uses the likelihood estimate. Table 7.2 shows that the default configuration was the weakest. The reduced set option, which limits the features to the five that are determined to be the best in terms of predictive power is very close to the default, but emphasizes recall over precision.

Two options produce big gains in performance: The kernel option uses a kernel function to estimate the density of continuous attributes, thus avoiding the assumption of normality in the distribution [JL95], and the discretization option uses an information theoretic metric, minimum description length, to group the attribute values and convert them to symbolic values. Yang and Webb [YW03] argue that discretization is effective when the probability density functions are not available, although they concede that the formation of the intervals

Configuration	Precision	Recall	Fmeasure
Default	0.640	0.875	0.739
Reduced Set	0.631	0.922	0.749
Kernel	0.662	0.921	0.770
Discretize	0.698	0.863	0.771

Table 7.2: Weka experiments comparing naïve Bayes performance. The reduced set uses only the five best features, which are determined by the implementation. The kernel estimator option uses a kernel for numeric attributes rather than a normal distribution, and the Discretization option converts numeric attributes to nominal ones.

---

are heuristic. In the Weka implementation, Table 7.2 show that discretization outperforms the other configurations, especially in the precision scores.

The question of what kinds of classification problems allow naïve Bayes to make the independence assumption safely has been pursued in the machine learning community[Ris01, Lew98, DP97].

If attributes are in fact independent of one another, then naïve Bayes is optimal, but this is rarely the case. Yet, the classifier is widely used and often "outperforms more powerful classifiers" [DP97]. It would seem that naïve Bayes performs well only when the entropy of the features of the instance space is extremely low, but it often works well under the opposite conditions. Figure 7.3 shows that the features in the **NIA** training corpus do not offer anything close to an optimal situation, in which there would be clear regions with one classification dominating the other.

Each of the six bar graphs were generated by Weka. Each bar along the x-axis shows a small range of values, determined by Weka. Note that these are not integral values since the words are weighted as described in Chapter 5. The light gray segments of the bars show the proportion of old instances, and the dark segments are the proportion of new instances. It's clear that none of these features is a good indicator of novelty by itself. The "Novel Word" graphs show the effect of previously unseen words. Where there are fewest, the leftmost bar, the proportion of old is clearly highest, and as one moves to the right, the proportion of

new rises, but the number of instances drop quickly. Each of the features depicted, and the others, as well, contain this kind of ambiguity. It underscores the importance of choosing the best method of combination.

7.1.2 Support Vector Machines

Support vector machines (SVMs) have become effective in many real-world classification tasks. Largely due to Vapnik [Vap95], they work by placing the training instances into a higher dimensional vector space, and then seek the optimal hyperplane separating the classes. The computation, via quadratic programming, is demanding. But various kernel functions have proven to be effective and more efficient. Not all of the training vectors are used, in an effort to improve the system's ability to generalize. The system seeks to find the transformed vectors that lie closest to the borders of the separating hyperplane.

SVMs avoid the local minima that hurt learning methods such as neural nets [Bur98]. They are largely characterized by the choice of kernel, which is an active area of research in the machine learning community. Common choices are a linear kernel, a polynomial kernel, radial basis function and a sigmoid. Tables 7.3 and 7.4 show a wide range of results depending on both the kernel and the implementation.

I tested two implementations and a variety of kernels available in both. Weka offers an efficient version of SVMs, using Sequential Minimal Optimization (SMO) [Pla99] (See Table 7.3), and SVM-Light [Joa98], which uses a different approach to speed up the optimization phase (See Table 7.4). When these two programs were tested on a variety of standard data sets in machine learning research, the results were mixed, with one having the advantage sometimes, and the other, the rest. Here, SVM-Light, with the RBF kernel, outperformed the SMO implementation. But Table 7.4 also shows that SVM-Light often converges on a hypothesis equivalent to the trivial baseline of "accept everything."

Interestingly, the SMO implementation did relatively well with both the polynomial kernel and the RBF kernel, when instance values were standardized, or transformed to a distribution with a zero mean and unit variance, as opposed to normalization, setting all values between $0 \ldots 1$. Another interesting point is that a rather large number of instances are support vectors in the SVM Light implementation, likely because of the noisiness of the

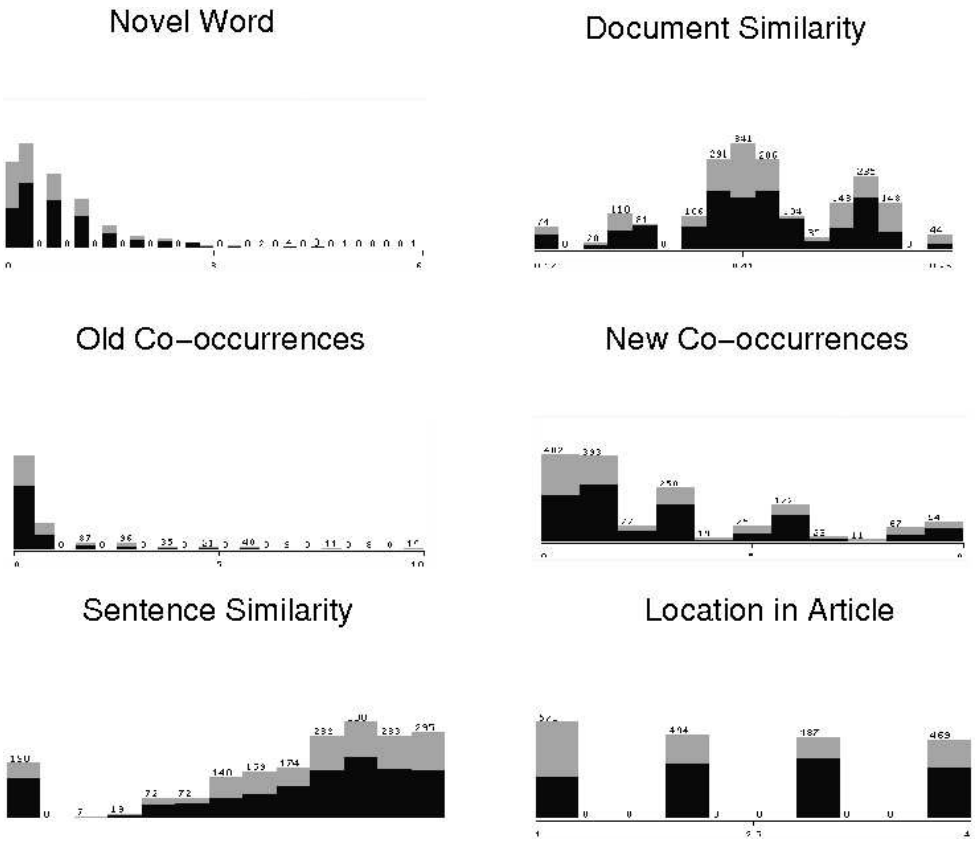

Figure 7.3: These charts portray the high entropy of the training cases. They show the breakdown of the classifications according to six representative features. Each bar along the x-axis represents a range of values for the feature. On the vertical axis the light and dark gray portions of the bars show how many cases are old and how many new, respectively. The fact that all the nearly are roughly even splits between old and new. The graphs were made by the data visualization facility in Weka. Low entropy would have resulted in regions where either the light or dark gray dominated.

SMO Parameters	precision	recall	F-meas
RBF kernel, normalize instances	0.605	1.000	0.754
RBF kernel, standardize instances	0.637	0.948	0.762
RBF kernel, no instance filtering	0.636	0.924	0.753
RBF kernel, standardize, gamma = 0.04	0.654	0.885	0.752
RBF kernel, standardize, gamma = 0.08	0.657	0.864	0.746
RBF kernel, standardize, gamma = 0.16	0.656	0.863	0.746
Poly kernel, defaults	0.637	0.934	0.758
Poly kernel, standardize, exp = 2.0	0.655	0.877	0.750
Poly kernel, standardize, exp = 3.0	0.665	0.877	0.750
Poly kernel, standardize, exp = 4.0	0.671	0.845	0.748
Poly kernel, standardize, exp = 5.0	0.668	0.844	0.746
Poly kernel, normalize, exp = 4.0	0.657	0.863	0.746
Poly kernel, standardize, complex = 7	0.665	0.739	0.700

Table 7.3: Experiments with Weka SMO version of efficient support vector machines. The implementation provides a choice of two kernels, the Radial Bias Function and the Polynomial. They have similar performance, although the polynomial kernels of degree > 2 have slightly better precision, at the expense of recall.

options	sv-count	precision	recall	F-meas
linear	1,600	0.605	1.000	0.754
poly,$d = 1$	1,618	0.605	1.000	0.754
poly,$d = 2$	1,623	0.605	1.000	0.754
rbf,$g = 0.1$	1,678	0.658	0.952	0.778
rbf,$g = 0.3$	1,734	0.671	0.940	0.783
rbf,$g = 0.5$	1,781	0.673	0.950	0.789
rbf,$g = 0.7$	1,799	0.670	0.961	0.790
rbf,$g = 0.9$	1,819	0.666	0.968	0.789
sig,$s = .5, r = .5$	823	0.605	1.000	0.754

Table 7.4: In contrast to the SMO implementation, SVM-Lite's RBF kernel outperformed the others. In fact the others all converged on the accept-everything solution.

---

data.

7.1.3 Instance-Based Learning

Instance-based learners classify inputs by comparison with annotated examples in its database. The K nearest neighbors algorithms are the basic form of this technique. Each input feature vector is classified by the labels of the k most similar annotated feature vectors.

The underlying idea is intuitively satisfying, namely that similar cases would have similar labels. Of course, the algorithm does not output a hypothesis for testing unseen examples; this feature pushes some of the work off to the online classification of new cases.

Weka implements a straightforward instance-based learner, the IBL system from [AKA91] and the more recent K* algorithm with an entropy-based similarity measure [CT95].

IBL uses a similarity metric using squared differences over the features for the n cases:

$$Sim(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)},$$

where $f = (x_i - y_i)^2$ for numeric attributes and $f = (x \neq y)$ for symbolic attributes. For each test instance, the classification function chooses the labeled instance y with the

K nearest neighbors	Precision	Recall	Fmeas
k=1	0.663	0.654	0.658
k=2	0.646	0.872	0.742
k=3	0.665	0.715	0.689
k=4	0.647	0.845	0.733
k=5	0.667	0.744	0.703
k=6	0.658	0.85	0.742
k=7	0.674	0.769	0.718

Table 7.5: Experiment with the Weka implementation of the IBL algorithm. The table shows precision and recall for runs where k was varied from 1 to 7. The results in terms of precision are similar to the other approaches, but recall tends to be a bit lower.

---

maximal value of f. In addition, the algorithm has a mechanism to record the accuracy of the k neighbors and can ignore the classifications suggested by those labeled instances that have poor records over the entire training set.

Table 7.5 shows a somewhat unexpected seesaw pattern as we increase the value of k.

The K* algorithm of [CT95] introduces an information theoretic metric for similarity between examples. The program computes a transformation of one instance x to another y. The distance between any pair of x and y would then be the shortest transformation necessary to turn one into the other, and the distance function used is:

$$K^*(y|x) = -\log_2 P^*(y|x),$$

The program takes one parameter, the blending parameter B, which in effect sets how many neighbors of the instance being classified are considered, where $B = 0$ is the algorithm behaves like a nearest neighbor algorithm and a value of 100 considers all training instances. There is also an option to set the blending parameter automatically according to the distribution of values for each feature.

Table 7.6 shows the results of varying the settings for B. As B was decreased toward the behavior of a nearest neighbor algorithm, precision steadily rose. When B was set to

	Precision	Recall	F-measure
Auto Blend	0.623	0.922	0.744
B=100	0.605	1.000	0.754
B=80	0.625	0.954	0.755
B=10	0.688	0.707	0.697
B=5	0.695	0.706	0.700
B=1	0.709	0.703	0.706
B=0	0.724	0.738	0.731

Table 7.6: K* version of instance based shows an improvement over IBL on a test using cross validation

higher values, precision remained close to the majority class, while recall was near perfect. As B approached 0, recall declined, but leveled off and even turned up at zero.

With instance-based learning, in particular the K* implementation, precision was finally pushed past 0.7. In the next two sections, the hierarchical characteristics of decision tree learning and rule induction will show even greater gains.

7.1.4 Decision Tree

Decision tree learning is accomplished by the construction of a tree structure which represent a set of rules induced from a set of classified instances. Each path to a leaf can be converted into a rule, or a conjunction of conditions, and the entire hypothesis is a disjunction of these conjunctions. The nodes of the tree represent the features in the problem space, and the arcs are constraints on the values for the feature represented by the parent node.

The basic learning procedure is to recursively choose the next best node for growing the tree. A typical way to make this choice is by an information theoretic notion of reducing entropy as much as possible. The entropy of a set of possible values W for some attribute is defined as:

$$H(W) = -\sum_{w \in W} p(w) \log(p(w))$$

Initially, the set W is the target classification, in our case, *new* or *old*. At each step we want to choose the attribute W_k that produces the largest information gain, that is the attribute with the smallest weighted sum of entropies over its values:

$$IG(S, W) = H(S) - \sum_{w \in W} \frac{|S_w|}{|S|} H(S_w)$$

S is the set of examples and W is the set of values for some attribute or feature in the hypothesis space. The next node with the largest gain is then chosen. Thus the search for a hypothesis is a greedy, top down search, always selecting the most favorable attribute at any particular point in the process.

A variety of methods have been developed to avoid overfitting in decision tree learning, from simple strategies of setting a minimum number of training examples to be covered by each rule, to the use of a validation set and statistical methods to weed out unlikely rules.

7.1.5 C4.5

C4.5 is an established decision tree program by Quinlan [Qui93]. It was developed during the 80s and early 90s and remains competitive with other learning systems. C4.5 builds in a mechanism to avoid the kind of fragmentation of training examples shown in Figure 7.2 that leads to overfitting and brittle performance. Information gain as shown above is normalized over the information in the outcomes at the nodes, what Quinlan calls split information:

$$split\ \ info(X) = -\sum_{i=1} \frac{|T_i|}{|T|} \log_2(\frac{|T_i|}{|T|},$$

for a test X on some attribute.

Several additional strategies are available in the Weka version of C4.5, which is called J48. The results are shown in Table 7.7.

- Reduced error pruning. Error rates for each leaf are computed on a subset of training instances that are held out when the initial model is built. The big drawback is the

need to hold back some of the training instances, which is especially serious for this work, since training instances are very expensive to obtain and limited in number. The initial trees are then built on fewer instances and may be less powerful to start with, and cross validation at this stage adds a considerable computational burden to the system.

- Minimum number of objects. This is a general tactic for rule-learning systems that requires leaves to cover a minimum number of cases. Rules that cover only a handful of training examples are suspect since they may perfectly fit rare combinations of values and fail to generalize. Under the 10-fold cross validation evaluation used in this work, the value of 7 worked the best. Table 7.7 shows it produced a substantial improvement over the default program options and over no pruning at all.

- C4.5 pruning. The default pruning method uses a probabilistic strategy that computes the likelihood of error at each leaf in the tree, but instead of using the exact count at the leaf, the program estimates a confidence interval based on the binomial distribution and uses the upper limit as the error figure. This value increases the error computation for nodes that cover only few cases and tends to eliminate them more frequently. The default confidence level is 0.25. Quinlan [Qui93] uses an example of a leaf covering six examples, all correct. So instead of using a 0 zero error prediction, he computes an error of 0.206. Smaller confidence levels lead to more aggressive pruning, and as Table 7.7 shows, *Conf .125* is beneficial to **NIA**.

- Laplacian smoothing. Another option, which was used in the early versions of C4.5, is to use Laplacian smoothing, adding a small value to all counts before computing the probabilities used in pruning. The value was 0.5.

7.1.6 Rule Induction

Rule induction proceeds much like the decision tree, conducting a greedy search through the problem space to develop rules covering the training cases. When the search is completed, the resulting hypothesis is pruned.

	Precision	Recall	F-measure
C4.5	0.714	0.787	0.749
C4.5 - red.err.	0.702	0.823	0.758
C4.5 - no prune	0.700	0.718	0.709
C4.5 - laP, minobj 7	0.718	0.812	0.762
C4.5 - laP, minobj 7, Conf .125	0.712	0.840	0.770

Table 7.7: Tweaking C4.5, as implemented in WEKA.

Ripper (Repeated Incremental Pruning to Produce Error Reduction) is a popular rule induction system that grows a hypothesis in much the way that a classic decision tree like C4.5 does [Coh95]. It proceeds by conducting a greedy search for rules that cover as many of the positive examples as possible, but rather than constructing a tree, it builds rules – a logical form composed of conjuncts.

Ripper uses a form of reduced error pruning, but rather than creating a complete hypothesis and then trimming back on rules that have high error rates on a held-out set of examples, it prunes as it goes along, for each new rule. The metric used for pruning is:

$$v(Rule, Pos, Neg) \equiv \frac{p - n}{p + n},$$

where Pos and Neg are the examples in the held-out set.

Rules continue to be added until the positive examples in the training set are all covered or until a threshold based on description length of the entire rule set is exceeded. In addition, there is an optimization option under which the rule set is examined again, rule by rule. For each rule, an alternate is constructed by growing and pruning against the error measured for the entire rule set. Table 7.8 shows that pruning of any kind raises precision to the best level seen so far, while maintaining a reasonably high recall. Without pruning, the results are similar to those obtained by support vector machines, i.e. the SVM Light version – very high recall but relatively low precision.

	Precision	Recall	F-measure
Ripper	0.721	0.834	0.773
Ripper No prune	0.666	0.944	0.781
Ripper Opt 5	0.723	0.823	0.770
Ripper Opt 10	0.727	0.810	0.766

Table 7.8: Tweaking Ripper, as implemented in WEKA, as JRip. The "opt" parameter controls the amount of optimization, which is done by post-pruning the rules.

7.1.7 Meta Learning

Meta learning is a diverse collection of ways to form a composite system of indvidual classifiers. Two of the best known approaches are *Bagging* [Bre96] and *Boosting*[Fre95, FS96], both of which combine multiple classifiers. Each method achieves significantly lower error rates than the underlying classifier, and each has been studied extensively in the past few years. We tested both these algorithms in the WEKA toolkit, in addition to a cost sensitive aggregation algorithm [Dom99], called MetaCost in WEKA. For all these experiments, the underlying learning algorithm was WEKA's version of Ripper.

7.1.7.1 Bagging

Bagging forms the multiple classifier by creating multiple training sets. For each of the desired number of classifiers, a new training set is creating by sampling the original training instances with replacement. This training set is normally the same size as the original, but in some of the replicated sets, a particular instance may not appear at all, and another may appear more than once.

The aggregate classifier assigns the class of each of the instances in the original training set by a majority vote.

A key condition for the system to work is an underlying instability of the underlying learner. If this condition is met, the changes in the training instances result in significantly different classifiers that generalize better.

7.1.7.2 Boosting

Boosting operates by forming different classifiers by assigning weights to each instance and adjusting these weights at each of the iterations building a new classifier. The idea is that the weights of the misclassified instances, that is the more difficult instances, are increased to focus the effort on them.

The error of each of subordinate classifiers is measured by the sum of weights of the misclassified instances. If the total error exceeds a threshold, or if it drops to zero, the iterations are stopped.

The aggregate classification is also formed by voting, as in bagging. The main condition for successful boosting is that the underlying classifier achieves 50% accuracy on the training set.

7.1.7.3 Cost Sensitive

Cost-sensitive learning addresses the problem highlighted by the use of precision and recall evaluation metrics here. Most classification algorithms are intended to minimize error rates without consideration of any cost differential. But in many real-world problems different types of errors have different consequences.

Weka's MetaCost algorithm is based on Domingos [Dom99], which wraps a meta-learning procedure around another classifier, and tries to reduce the *risk* associated with labeling instance x as a member of class i. It is computed as:

$$Risk(i|x) = \sum_j P(j|x)Cost(i,j),$$

where $P(j|x)$ is the likelihood of an optimal classifier assigning label j to the instance x. MetaCost is related to bagging in which multiple models are built on several samplings of the training instances. MetaCost also learns multiple models, but estimates the class probabilities $P(j|x)$ by the number of votes that class receives from the different models and then relabels the training examples, so that the new labeling reflects the differential in the cost matrix, $C_{i,j}$. The success of the MetaCost algorithm suggests further experimentation along this line, in particular the approach by Fan and others [FSZC99], which adds a

	Precision	Recall	F-measure
AdaBoost	0.735	0.814	0.772
Bagging	0.725	0.892	0.800
MetaCost	0.762	0.729	0.745

Table 7.9: Varieties of meta learning: Boosting, Bagging and Cost Sensitive learning

cost factor to the weighing computation for the multiple hypotheses created in Boosting algorithms.

7.2 Conclusion

The experiments in this chapter show that learning algorithms can push precision to more useful levels, without reducing recall. In doing so, it was clear that rule learners were the best suited for the complicated data. Both the Weka versions of Ripper and C4.5 are able to achieve interesting partitions of the data on the basis of the proposed feature values. At the end of this series of experiments, Ripper seemed like the best choice as the base algorithm. Its results were the strongest, especially with respect to precision, and it is efficient enough to be used in the metalearning algorithms, which offer further improvements. Table 7.9 shows that boosting, Weka's implementation of AdaBoost, raised precision above the best results Ripper had obtained alone, without harming recall. Bagging, in Weka, matched Ripper's best precision, and raised recall to a very high level. Finally, Weka's MetaCost algorithm, setting the penalty for false positives at twice the value of that for false negatives, produced a substantial gain in precision. In addition, a rule learner like Ripper produces understandable rules that can be used to examine the plausibility of the results.

In the next chapter, Ripper will be used to conduct a series of experiments to determine the strongest options for running **NIA**.

Chapter 8

Evaluation

The evaluation in this thesis concentrates on establishing the overall system's ability to determine which input passages contain novel information. Now, that a learning algorithm has been chosen, i.e. Ripper, there are many choices in how to run **NIA**; options such as which lexicon and method of forming co-occurrences need to be tested in order to determine how best to use the system.

I use a black box evaluation, testing the system as a whole, without looking into the performance of any of the components. Indeed, the subsystems were built specifically for **NIA** and it is not clear how to measure them in some other context. Furthermore, initial processing of the input text is mainly performed by two off-the-shelf programs, Talent [RWC97], a named-entity recognizer, and a parser [Cha00] and evaluation of them as stand-alone systems is outside the scope of this work. The basic operation of our system is to take these inputs, along with semantic and lexical information, and to extract and combine different sets of features. The features are the input to the learning programs, and again, I am using an off-the-shelf suite [WF00]. One contribution of my research is in the particular selection and extraction of features and the way they are integrated.

Thus the combination of various strategies and components will be examined, by adjusting and changing the system's options, and then by examining the results when learning over partial sets of features. The experiments will encompass the following:

- Choice of features from those in Chapter 7.

- Choice of semantic dictionary. There are six variations.
- Use of document frequency and word-promiscuity to measure word content, as described in Chapter 5.
- Choice of words, i.e. whether to use head words only.

The primary evaluation will be against the development set of 31 pairs of news articles described in Chapter 3, using *JRip*, the WEKA version of Ripper [Coh95]. I use 10-fold cross validation for the evaluation in order to make the best use of the 2,023 clauses in the annotated corpus. The classification of each was agreed upon by two annotators (See Chapter 3).

In addition, I will test the system on the much larger Text Retrieval Conference (TREC) 2004 Novelty Track corpus, which consists of 8,343 sentences[1] These sentences were drawn from some 2,500 news articles grouped into 50 clusters, each centered on a given topic. There were a total of 52,447 sentences in the clusters. The pure novelty part of the Track was to scan the 8,343 relevant sentences in order and filter out any that simply repeated information contained in the previous sentences. Of the 8,343 relevant sentences, only 3,454 were classified as novel – or 41.3%.

There is a mismatch between the task defined for **NIA** and that of the Novelty Track. My system has features that are based on units smaller than a sentence and other features that express an elementary notion of context. On both counts, the Novelty track is different. It is based solely on sentences, and the sentences are removed from their original context. Nonetheless, these trials were informative.

[1]There was considerable difference of opinion between the two NIST assessors. NIST did not calculate a measure of agreement, like the Kappa co-efficient; however, the two assessors disagreed on at least 3,410 sentences. This figure is derived by assuming that the intersection of selections by both assessors was maximal, and so that the disagreement is computed by subtracting the number of selections by one assessor from the other.

8.1 Choice of Features

The system is capable of generating a large set of features, with one group of metrics based on the co-occurrences from the *Micro View*, and another based on context and vocabulary from the *Macro View* from the SumSeg system (See Chapter 7). I added one feature measuring sentence similarity, following the baseline vector-space/cosine system I tested in the Novelty Track of TREC 2004. A last group of features describe surface structure, like size of the unit and location. Finally, I tried a feature measuring document similarity, which initially seemed to produce a big boost in results, but which was problematic on closer inspection, as shown below. For the discussion of the different features in this section, I exclude the *docsim* feature from this comparison. The comparisons were done with all 31 pairs of documents in the corpus annotated as described in Chapter 3.

To compare various features, I examine the behavior of the rule learner, the WEKA version of Ripper [Coh95], using 10-fold cross validation. I chose a configuration of program options that scored very high in preliminary rounds of tests – using the very basic lexicon of WordNet synonyms only, without the promiscuity lexicons (See Chapter 5), and using all nouns and verbs in the units, leaving out adjectives. No penalties were imposed for document frequencies or promiscuous words (See Chapter 5). I held this configuration constant for the test of features, and successively dropped different groups of them to observe how they behave.

In testing, the combination of all the features by far does the best, in terms of both precision and recall. Table 8.1 shows the results of the different feature subsets in isolation and in combination. All the results of subsets are below those of the full combination. The difference between the full combination and partial combinations of two or three groups is not statistically significant, but the full combination is statistically significantly better (with $p < 0.01$) than any of the individual subsets of features. When each group is isolated, the differences are much clearer. This result is important because it shows the value of the new approaches – from **NIA** and SumSeg themselves – over the more standard sentence by sentence comparison. It also show the value of them individually. The **NIA** features produce the highest precision, while maintaining a reasonable recall. The SumSeg features are in between, and the sentence-based system, Cosine, produces a degenerate solution,

Features	Correct	False Pos.	False Neg.	Precision	Recall
All	984	413	239	0.704	0.805
NIA, Sumseg, Struct	979	430	244	0.695	.800
Sumseg, Cosine, Struct	956	422	267	0.694	0.782
NIA, Cosine, Struct	977	428	246	0.695	0.799
NIA and Struct	961	424	262	0.694	0.786
SumSeg and Struct	969	441	254	0.687	0.792
Cosine and Struct	970	462	253	0.677	0.793
NIA alone	933	467	290	0.666	0.763
SumSeg alone	1056	600	167	0.638	0.863
Cosine alone	1223	800	0	0.605	1.000

Table 8.1: A comparison of the system with only partial sets of the features. When all the features are combined, the results are the best of any of the subsets, and significantly better than any of the single subsets (using the binomial test for total correct classifications).

---

always choosing the majority class, i.e. accepting all the units as novel.

In terms of precision, which is clearly the harder part of the task, the **NIA** subset performed the best, with a statistically significant advantage over the SumSeg and Cosine feature subsets. In terms of an F-measure combination of precision and recall, the Cosine system does the best, but the learner settles on a degenerate hypothesis of always selecting the majority class, which is *novel* and therefore always gets a perfect recall. Such a system is useless in practical terms. The F-measure with equal weight to precision and recall is not a particularly good evaluation metric in this task because the same score can describe vastly different systems, and artificially high scores can be obtained whenever there is a large portion of positive cases – as in novelty detection (See Chapter 6). The F-measure provides a parameter, β, with which the emphasis can be shifted to either precision or recall, but normally the metric is used with β set to 0.5, giving both equal weight. The computation is as follows:

Feature Set	$\beta = 0.5$	$\beta = 0.61$	$\beta = 0.7$
Cosine	0.753	0.715	0.685
SumSeg	0.733	0.710	0.692
NIA	0.710	0.701	0.692

Table 8.2: Adjusting the β parameter in F-measure shifts emphasis to precision instead of recall, and evens out the F-scores produces by the three feature subsets.

$$F = \frac{1}{\frac{\beta}{Prec} + \frac{1-\beta}{Recall}}$$

Table 8.2 shows how small shifts in the β value of the F-measure evens out the performance of the three feature subsets. At $\beta = 0.5$, the accept-all behavior learned with using cosine similarities has the highest F-measure. The difference in F-measure shrinks to a few points when $\beta = 0.61$, a value selected because it reflects the split in the data, and disappears at $\beta = 0.7$, in which precision is favored. These results show that the **NIA** features are effective when high precision is desired and a large degree of compression is wanted to produce a smaller amount of output.

8.1.1 A Problem Feature

The document-similarity feature was the last to be added and resulted in a significant jump in both precision and recall. The intuition was that the comparison of passages might be measured with one set of features when the parent documents were similar and another when they were farther apart.

For example, Figure 8.1 shows four rules produced by Ripper. Those labeled A, B and C, which are plausible, in that they can generalize. Rule A says that a clause in the first quarter of an article that has a high docsim value should be labeled as *old*. Rule B say that a clause unit that contains no new novel words and has a high docsim value should be labeled as *old*. Rule C says that a unit with a low docim value and a low average set weight for its novel Co-occurrences should be classified as *old*. Average set weight is a normalized factor based on the frequency of the term in the document set under consideration, based

A: quad ≤ 1, docsim $\geq 0.571573 \rightarrow$ OLD 67/9
B: novel ≤ 0, docsim $\geq 0.571573 \rightarrow$ OLD 28/1
C: docsim ≤ 0.265878, wsmean $\leq 0.000004 \rightarrow$ OLD 41/13
D: docsim ≥ 0.431096, docsim $\leq 0.431178 \rightarrow$ OLD 100/25

Figure 8.1: A sample of four rules that reference the docsim feature – the highest similarity of the current document to any previous document. Rules A through C are plausible in that they can be expected to generalize. Rule D is not, basically choosing two mid-range values and declaring that labeling all of those clausal units adds the most to the score. The numbers after the classification show the data split as *correct*/*incorrect*.

---

on the assumption that the key terms in a set are the most often repeated. But Rule D is worrisome. Rule D says that a unit with *docsim* 0.431 should be classified as *old*. This value is the midrange for the docsim feature. It would in effect cover all units in two documents as *old* because those documents in this pair happen to contain little new material. At the same time, an examination of the input documents shows 9 other document pairs have $docsim = 0.431 \pm 0.05$ that are not covered.

A large part of the improvement from the docsim feature appears to be random, due to the chance circumstance that one of our 31 document pairs had very few novel passages in it. Although the learning algorithms include mechanisms to prevent overfitting, by pruning rules that cover too few examples, they fail to catch the kind of overfitting that applies to the docsim feature. They fail because all the units in a particular document share the same feature value. Rules like D in Table 8.1 appear in most configurations. The fault lies with the layout of the training data. The 2,023 units in the training corpus are sufficient examples to learn over (See Section 8.5) for clausal unit features, but there are only 31 document pairs. It is interesting that the learner covered both document pairs. Figure 8.2 also shows that novel units in the newer article of a pair of documents are not correlated to the degree of similarity between the two documents, making any rule like Rule D likely to be brittle.

I tried grouping the docsim feature values into a small number of bins, but found that the learner either ignored the feature when the number of bins was too small, or else it

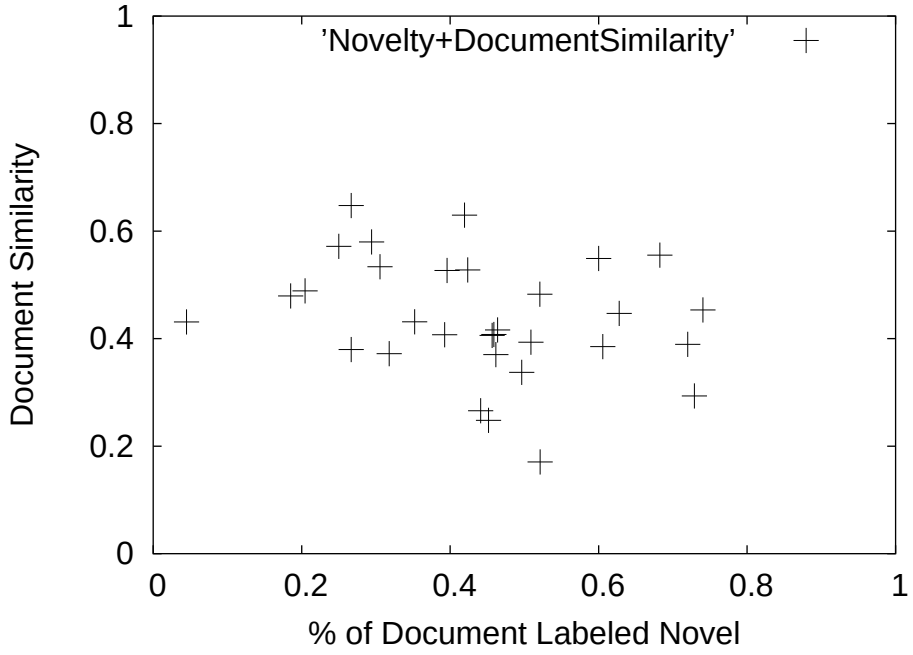

Figure 8.2: The plot shows the relation between the percentage of novel sentences in a document and its similarity to the earlier document. These values were calculated for the 31 pairs in the corpus described in Chapter 3. It is apparent that there is no strong correlation between the two, suggesting that the docsim feature is not by itself evidence of novelty.

---

came up with a very similar rule when the number of bins was too large.

The fact is that the docsim values over these 31 pairs do not correlate well with the proportion of novelty in the pairs (See Table 8.3).

As expected, removing the docsim feature cuts into the learning algorithm's performance, from 2.2% to 4.4%, depending on the various program options. Figure 8.3 shows the consistent difference between including the docsim feature and removing it.

Further experiments suggest that the docsim feature is valuable. I removed the two skewed documents from the training set and reran Weka's Ripper on two runs. Table 8.4 shows that the docsim feature still adds some discriminatory power to the learning algorithm on the reduced set of training cases. It's interesting that the results are better without

Pearson's cor	-0.258
Kendall's tau	-0.202
Spearman rho	-0.306

Table 8.3: While there is the expected negative correlation between the amount of novelty in the new document and the similarity of the new and old documents, it is too weak to rule out chance.

	Precision		Recall	
	Docsim On	Docsim Off	Docsim On	Docsim Off
	Reduced Training Set			
All Words	0.743	0.738	0.842	0.842
Heads Only	0.734	0.720	0.843	0.830
	Full Training Set			
All Words	0.729	0.706	0.837	0.818
Heads Only	0.731	0.688	0.764	0.716

Table 8.4: The tables shows the effect of including or removing the docsim feature from a test on a reduced training set. The two documents that resulted in apparently brittle rules were removed to make the reduced set.

the two documents that were removed. It may be that one or both of those articles was particularly difficult or unusual, but an exploration will have to be left for future work, especially since it would require additional training material.

8.2 Semantic Lexicons

I constructed six semantic lexicons to be used as plugins to **NIA**. These are used to find referential equivalents, as outlined in Chapter 5. The first is taken from Dekan Lin's semantic dictionary [Lin98a], an automatically constructed dictionary based on patterns of syntactic usage. The second is an empty lexicon that simply provides the uninflected forms of the

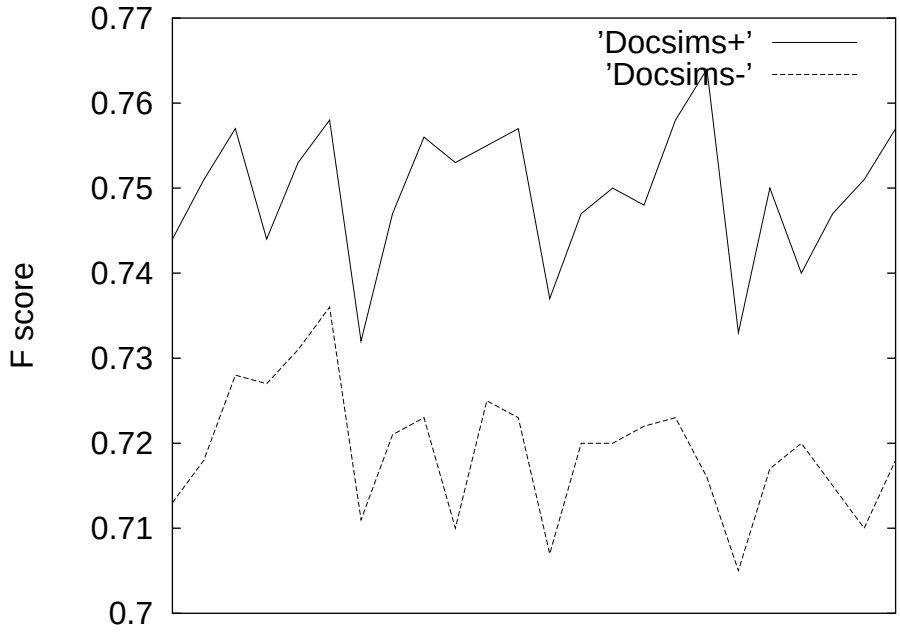

Figure 8.3: The graph shows the effect of the docsim feature. The results from 24 different program configurations are shown along the horizontal axis. The line on the top gives the results with the docsim feature, and the one on the bottom, without. The scores are the F-measure for the different runs.

---

words. Three of the remaining variations are taken from WordNet[MBF+90]. The WordNet database is converted into a convenient form for the program, and is easily altered to allow for additions from other sources, currently one from NOMLEX[MGM+98] and one NOMLEX and CELEX[CEL95]. I also used a combination of Lin's dictionary with WordNet. For this chapter, the lexicons are labeled as follows:

Dekan The lexicon taken from Dekan Lin's work.

Empty The empty lexicon.

Minimal WordNet synonyms.

Combine The intersection of all Minimal and Dekan.

Nominals Minimal plus immediate hypernyms and hyponyms plus Nomlex.

Morph Nominals plus morphological transformations from Celex.

I tried all six lexicons on the 31 pairs of articles in the development set, varying **NIA**'s options. For these experiments, I used the same 0.7-precision/0.3-recall weighted f-measure as in Subsection 8.1.1. In general, I found rather small significant differences among the lexicons. The one consistent result was that the automatically derived dictionary, *Dekan*, underperformed the others. Inspection shows that many entries are quite noisy. In all, I tested the six lexicons on 12 different sets of program options, and then compared their results against each other. In these 12 experiments, the *Dekan* lexicon scored the lowest in 9, and the *Combine* lexicon was the lowest in the remaining 3. The *Morph* lexicon, which is the most sophisticated of the six lexicons, was the top scorer in 6 experiments. (See Table 8.5, and in Appendix D, Tables D.2, 8.6, D.7, D.8 and D.9).

Lexicon	Precision	Recall	F-measure
Dekan	0.691	0.772	0.713457268958138
Empty	0.697	0.773	0.718183151159691
Minimal	0.700	0.805	0.728506787330317
Combine	0.703	0.792	0.727526460211682
Nominals	**0.704**	**0.805**	**0.731534787659739**
Morph	**0.706**	**0.818**	**0.736241713411525**

Table 8.5: All words, no promiscuity or document-frequency weighting and no docsim feature. The two rows in bold are in the top five performers for all configurations without the docsim feature.

Table 8.5 is a good example of the differences between lexicons. For this table, **NIA** used all the words in the clause units to form Co-occurrences, used no weighting for promiscuity or word document-frequency, and no docsim feature. The rows in bold face indicate that these were among the top five results in all the experiments without the docsim feature.

The same pattern holds when the results with the docsim feature are included. Table 8.6 shows the numbers are higher but the **Morph** lexicon remains the best, although in this experiment, the **Minimal** lexicon scores very high, especially in precision.

Lexicon	Precision	Recall	F-measure
Dekan	0.715	0.820	0.743563728598605
Empty	0.725	0.818	0.75060118972282
Minimal	**0.739**	**0.803**	**0.757102577188058**
Combine	0.719	0.811	0.744331120755681
Nominals	0.728	0.817	0.7525952170062
Morph	**0.729**	**0.837**	**0.758355704697987**

Table 8.6: All words, no promiscuity or document-frequency weighting but this time with the docsim feature. The two rows in bold are in the top five performers for all configurations with the docsim feature.

Throughout this second round of experiments, the *Minimal* lexicon did well and was the top scorer 4 times (See Table 8.8, and in Appendix D, Tables D.4, D.5, D.6). The *Nominals* lexicon was the top scorer twice (Tables D.1 and D.3). The empty lexicon was never either the best nor the worst in these experiments. These tested not only the choice of lexicon, but whether or not to include adjectives, whether or not to restrict the Co-occurrences to heads of phrases, and whether or not to use some kind of content weighting, either promiscuity, document frequency or both.

Over all, it seems like the application of semantic information helps, but only to a limited amount. It is difficult to isolate the effect of using one lexicon as opposed to another because of the interaction with the choices of other program options. In order to measure the differences among the dictionaries, I used the two-sided, paired t-test to determine if the scores over all 12 option combinations that I tested differed significantly from one another. In other words, I wanted to see if the different results from using different lexicons could have been random variations in imposing constraints on the formation of co-occurrences at run-time.

Table 8.7 shows the results of the paired t-test on the f-scores of these 12 experiments. We can see that the automatically extracted *Dekan* lexicon performed below all of the WordNet-based lexicons (*Minimal*, *Nominals* and *Morph*) with a $pvalue \leq 0.1$, and below

	deklin	empty	minimal	combo	nominal	morph
deklin	-	$p \leq 0.01$	$p \leq 0.01$	$p \leq 0.05$	$p \leq 0.01$	$p \leq 0.01$
empty		-	$p \leq 0.05$	NULL	$p \leq 0.05$	$p \leq 0.01$
minimal			-	$p \leq 0.05$	NULL	NULL
combo				-	NULL	$p \leq 0.01$
nominal					-	$p \leq 0.05$
morph						-

Table 8.7: Results of the two-sided paired t-test on 12 different tests of the six semantic lexicons. The values compared are the $\beta = 0.7$ f-measure, weighted to give slight emphasis to precision scores.

---

the *Empty* lexicon. The differences between *Empty* and the WordNet lexicons were not so clearcut. The *pvalue* in comparing *Empty* with either *Minimal* or *Nominals* is ≤ 0.05. The *Combine* lexicon was interesting since it took the intersection of an expanded selection of related words from WordNet and the *Dekan* automatically acquired lexicon. In most of the configurations, *Combine* held a middle rank, but it varied quite a bit, being the low scorer in the configuration using all the words except adjectives without any promiscuity penalties (Table 8.8), and in the configuration using all the words, but with document frequency as the only determinant for low content (Table D.6).

Of the top five results (bold typeface in Tables 8.5 and 8.8) in the experiments without the docsim feature, two were produced with the Morph lexicon, two with the Nominals lexicon and one with the Minimal lexicon. The results were extremely close to one another, with precision ranging from 0.704 to 0.714, and recall from 0.799 to 0.825. These are not significantly different from one another, either from the perspective of precision or recall, according to the binomial test.

Tables 8.6, D.8 and D.9 show the five top scores in the experiments with the *docsim* feature turned on. Three of these were with the Morph lexicon and one each with the Nominals and the Minimal. Like the results without the *docsim* feature, these are not significantly different from one another.

1. quad ≤ 1, cover $\leq 0.775934 \rightarrow$ old (332.0/114.0)
2. bigrms ≥ 1, glodist $\geq 3.872983 \rightarrow$ old (246.0/103.0)
3. meanset ≥ 0.000143, cover $\leq 0.69678 \rightarrow$ old (74.0/31.0)
4. otherwise $\rightarrow$ new (1371.0/396.0)

Figure 8.4: One of the simplest yet most effective rulesets, learned from the configurations, Table 8.8, using the *Nominals* lexicon, without the *docsim* feature and without promiscuity, considering all nouns and verbs, but no adjectives. Average performance on cross validation was equal to the hypothesis formed by the learner over the full set of training instances. The rules also avoid taking narrow slices of feature values as dubious rules that might not generalize.

Figure 8.4 shows the rules generated with the *Nominals* lexicon from Table 8.8. When the system is classifying new, unseen, unlabeled instances, these rules are applied one by one, in order. If any of the first three fire, the classification is done and the classifier proceeds to the next rule.

This configuration is one of the most successful achieved without the *docsim* features, posting a precision-biased f-score of 0.738. Its precision of 0.714 is the highest of all experiments without the *docsim* attribute. Even more telling is the fact that the ruleset seems to generalize well, with the average in the 10-fold cross-validation equal to the performance of the training data. The ruleset is small and avoids reliance on taking narrow slices of values from any of attributes. The rules also represent a combination of **NIA**'s Micro View, SumSeg's Macro View and the sentence-based vector-space strategy. Rule 1 combines a structural feature, the *quad* feature, which is the location of the unit in the article, with the sentence similarity computation. Rule 2 combines the *bigrams* feature, a Micro View feature that is a weighted sum of the old co-occurrences in a clause, and the *glodist* feature, a Macro View feature that measures the distance between the current clause and the last clause that contained a novel Co-occurrence. Rule 3 is another mixture of the different approaches, combining a Micro View feature, the *meanset* feature, which is the average weight of the novel Co-occurrences, with a measure of sentence similarity.

This kind of rule is just what I sought from the application of machine learning to the problem. It mixes features derived from different perspectives, and achieves a substantial increase in precision without much loss of recall. The numbers in parentheses alongside each rule in Figure 8.4 show the classifications and error rates on the model built from the full training set. These amount to precision of 0.711 and recall of 0.797, which are very close to the 10-fold cross validation averages in Table 8.8. The result also shows a distinct benefit from using the semantic lexicons that are based on WordNet.

Dekan	0.703	0.765	0.720518488745981
Empty	0.692	0.798	0.720720438527799
Minimal	**0.711**	**0.825**	**0.741748861911988**
Combine	0.682	0.774	0.707218649517685
Nominals	**0.714**	**0.799**	**0.737538461538462**
Morph	**0.708**	**0.810**	**0.735796766743649**

Table 8.8: No adjectives, no promiscuity, no docsims

8.3 Promiscuity

My goal here was to weight common nouns and verbs according to a measure of their content. I applied the two metrics described in Chapter 5, the word *promiscuity* metric and a document frequency measure in order to eliminate erroneous Co-occurrences due to polysemy and generality. Linguists talk of empty nouns, that is, nouns that carry no content on their own, that behave like pronouns, acting merely as referents to some other entity. I sought to extend this idea to nouns that have little content. When building the lexicons, I imposed a threshold for the number of WordNet senses that a word can have, but this is not a reliable test because of the way WordNet is built. Often very closely related senses are separated, while some words like *object* are very general or vague but have few senses in WordNet. With a metric of word promiscuity, vague or general Co-occurrences would be devalued and improve accuracy.

Table 8.9 shows that the idea as implemented does not work. Rather than improve

Configuration	Without Promiscuity	With Promiscuity	% change
All words, no *docsim*	0.726	0.719	-1.0%
All nouns, verbs, no *docsim*	0.728	0.718	-1.4%
Heads only, no *docsim*	0.718	0.714	-0.6%
All words, with *docsim*	0.751	0.750	-0.1%
Heads only, with *docsim*	0.751	0.746	-0.7%

Table 8.9: A comparison of configurations with and without promiscuity. The changes are small, and they are not statistically significant, but the addition of promiscuity consistently brings the results down.

---

results, the addition of promiscuity values to the program leads to some deterioration. Although the differences are not statistically significant, according to the paired t-test, the results with the promiscuity metrics are consistently lower than those without.

I took a closer look at the outcome of the learning algorithm on comparable tests with the identical configuration of program options except for promiscuity weighting. This configuration, using the Morph lexicon and not including adjectives had relatively strong results over all. It is clear that the rules induced by the learner are quite different. Figure 8.5 shows the hypothesis from the no-promiscuity test. It is a mixture of Micro and Macro views, with three rules involving the *Cover* feature – based on the sentence-based vector-space approach. Rule No. 5 seems to be the type of brittle rule discussed in Section 8.1.1 of this chapter. Figure 8.6 shows the corresponding rules from the run with promiscuity. The rule set is much simpler, and cleaner, but its performance lags quite a bit. It relies heavily on the structural *quad* feature, which shows the location in the article (with each article divided into four quarters).

Thus, Rule No 1 says to classify everything in the top one-fourth of each article as old. Intuitively such a rule seem arbitrary and somewhat brittle. This suggests that more work is needed to see if the promiscuity has value despite the somewhat lower bottom-line result.

1. quad ≤ 1, bigrms $\geq 1 \rightarrow$ old (250.0/75.0)
2. nouns ≤ 0, cover ≤ 0.759635, quad $\leq 2 \rightarrow$ old (190.0/69.0)
3. novel ≤ 0, sntdist $\geq 1 \rightarrow$ old (55.0/12.0)
4. bigrms ≥ 1, quad ≥ 4, cover $\leq 0.859886 \rightarrow$ old (77.0/28.0)
5. cover ≥ 0.674205, cover $\leq 0.687224 \rightarrow$ old (30.0/7.0)
6. otherwise $\rightarrow$ new (1421.0/389.0)

Figure 8.5: The set of rules for the *Morph* lexicon in the configuration using all nouns and all verbs, but no adjectives, without promiscuity (See Table 8.8). It showed relatively strong performance and has a reasonable ruleset. It outperformed its counterpart with promiscuity.

---

1. quad $\leq 1 \rightarrow$ old (573.0/244.0)
2. bigrms ≥ 2, novel $\leq 0 \rightarrow$ old (67.0/20.0)
3. otherwise $\rightarrow$ new (1383.0/424.0)

Figure 8.6: The set of rules for the *Morph* lexicon in the configuration using all nouns and all verbs, but no adjectives, with promiscuity controls on (See Table D.2).

8.4 Choice of Words

One of the main program options defines what types of words in each unit are used in computing the various metrics for the features. We tested using all the content words, all nouns and verbs, excluding adjectives, heads of all phrases in each unit, and heads of major constituents.

When testing the configurations without promiscuity and without using the *docsim* feature, there is a clear advantage to using all the words, or all nouns and verbs, excluding adjectives, over the configurations using heads. The average f-measure for the all words is 0.726 over the six lexicons we tested; the average for excluding adjectives is 0.728. Meanwhile the average for the heads only configuration is 0.718. The paired t-test shows this difference is significant with $pvalue \leq 0.05$.

This distinction is lost if we add back either the promiscuity information or the *docsim* feature. With promiscuity turned on, the f-score for all words averages 0.719, and all but adjectives, 0.718. This is only slightly better than using only heads, where the average is 0.714. The difference is not statistically significant. When we include the *docsim* feature, there is no difference without promiscuity – with both all words and only heads both average 0.751. If promiscuity is added back in, there is a slight advantage to using all words over heads only, 0.750 compared with 0.746, but the difference in not statistically significant.

8.5 Amount of Data

One issue about the learning algorithm remains to be addressed. In order to check that the training corpus was sufficiently large, I divided it into training and testing portions, and then took drew randomly selected smaller training sets, starting at size 100, to examine the learning rate. For this experiment, I removed the examples from the skewed documents so that I could use the docsim feature and at the same time avoid the possibility of artificially inflating the result. Figure 8.7 shows that the learning rate rises quickly and starts to level off at just a few hundred annotated examples. This experiment was done with a subset of the examples, excluding two skewed documents that Ripper took advantage of. Thus, I conclude that I have a sufficient number of training cases with 2,023 for my experiments,

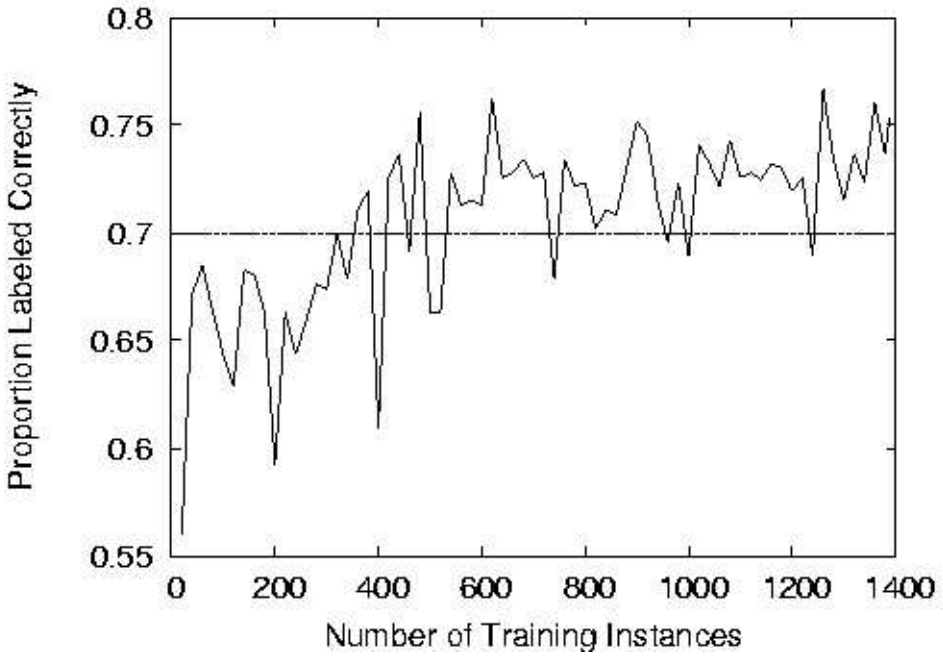

Figure 8.7: After randomly dividing the annotated instances 75% to 25% into training and testing sets, the training sets were further divided into successively larger sets to examine the learning rate and check that a sufficient number of examples were annotated. Each of the smaller training samples were chosen at random from the training portion at intervals of 20. For this test, the overall error rate was used.

---

when improvement starts to level off at 400 examples.

8.6 TREC

In Chapter 6 we presented results of a purely *Macro View* system, the exploratory SumSeg system, at the TREC 2004 Novelty Track. Its performance was strong compared with the other submissions, especially with respect to precision, which seemed to be far more difficult to achieve than recall in the Novelty Track evaluation.

The TREC Novelty Track guidelines do not specifically address how the assessors should handle sentences that are partially new and partially old. For the **NIA** development corpus, the instructions to the annotators were to mark only those words which encompassed new

information. In some sense, any choice of an evaluation unit will create some kind of imbalance. By adopting clauses as the unit, the evaluation might be skewed toward larger segments, even dominated by them, since a large cohesive segment of text would contain numerous clauses. But if a metric based on a larger unit had been used, changes about specific facts and events would have been lost. For example, a segment of m words with a novel phrase of n words. A ratio of $\frac{n}{m}$ would have to be specified that separates segments into novel and not novel parts.

To take advantage of the large amount of data available from the TREC 2004 Novelty Track, and to try **NIA** on larger sets, I adapted the program to operate on sentences. Table 8.10 shows that the choice of features is less important than the choice of learning algorithm. Naïve Bayes consistently achieves higher recall, while Ripper, a rule induction method similar to decision tree learning, tends to get higher precision. In both cases, the system obtained higher precision scores than any TREC system except for the Precision oriented runs by SumSeg, which were described in Chapter 6. However, many of the competitors at the TREC evaluation had much better recall scores, including the SumSeg runs. One possible reason is that the Micro View features of **NIA** are based on an analysis of subsentential units and were not designed to deal with whole sentences, yet the Novelty Track judgments are made only on a sentence basis. Another reason is that the features are extracted statically, off line, separate from the learning algorithm. In our Novelty Track system, the classification of a sentence, S_n, is made at run-time and depends on the decision made for sentence S_{n-1}. Standard learning algorithms require that training examples be preclassified. This would mean that we would need an exponential number of alternate classifications for each instance. But the experiment suggests that a more elaborate way of computing context would benefit **NIA**.

Meanwhile, Table 8.10 shows virtually no difference between the different feature sets. The tests shown in the table are the results from using 10-fold cross validation, and the F-measure was balanced equaled between precision and recall, in keeping with the TREC policy. In order to examine the errors more closely, I ran some tests by dividing the sentences into training and testing sets, random selections of two-thirds and one-third of the relevant sentences.

Run ID	Precision	Recall	F-measure
Macro View Only for TREC			
Naïve Bayes	0.53	0.62	0.570
Ripper	0.56	0.49	0.527
Macro View ∩ Cosine			
Naïve Bayes	0.53	0.62	0.570
Ripper	0.56	0.50	0.527
Micro View, with Structural			
Naïve Bayes	0.52	0.67	0.582
Ripper	0.57	0.47	0.517
All But Document Similarity			
Naïve Bayes	0.52	0.67	0.582
Ripper	0.56	0.48	0.519
All Features			
Naïve Bayes	0.53	0.67	0.591
Ripper	0.58	0.49	0.530
All Features, No Promiscuity			
Naïve Bayes	0.0.53	0.67	0.594
Ripper	0.57	0.500	0.532

Table 8.10: Results of experiments with on the TREC 2004 Novelty Track data after adapting **NIA** to operate on sentences.

I found that there is a substantial difference in the classifications when different feature sets are used. For example, consider the classifications made by Ripper with all the features and Ripper with only the Novelty Track features. There were 924 and 934 errors, respectively – out of 2,562 test cases. These include both false positives and false negatives. Yet the same errors were made by both feature sets on only 805 cases, suggesting that some improvement in extracting the features could translate into a substantial gain in performance.

8.7 Conclusion

The overall conclusion from the experiments in this chapter complement and amplify the results shown in Chapter 7: machine learning is effective in attacking the difficult problem of new-information detection. Chapter 7 suggested that the Ripper [Coh95] rule induction algorithm is appropriate for this task. With it, the experiments here show that it is the combination of groups of features that work the best. Section 8.1 tests each group separately, Micro View features, Macro View features and structural features. When isolated, the Micro View features work the best, bolstering the case for using clauses over sentences (Chapter 4. In addition, small gains are achieved by using surface semantic information from WordNet and other manually build resources (Chapter 5), but more work is needed to apply automatically obtained data such as Dekang Lin's lexicon [Lin98a]. Likewise, more work is needed for computing the importance of words.

Chapter 9

Summarization

In this chapter I will discuss how **NIA** is be used in a real-world application. The application is Columbia's NewsBlaster, a news browsing system. NewsBlaster will provide the clustered input documents to **NIA**, which outputs an exhaustive list of sentences – all those that contain a clause judged to be new – in the order of the input sentences. Because this output is often larger than what NewsBlaster expects, the last stage is to run it through a summarizer. At present, I use the DEMS [SNM02], which is described below in Section 9.1.

Every day, NewsBlaster crawls about two dozen news-dedicated websites such as The New York Times, The Washington Post and CNN, and downloads thousands of pages. News articles are automatically identified and their content is extracted. The articles on the events of the day are clustered and those on the same topic are placed together. In all, fewer than 1,000 are eventually summarized.

NewsBlaster offers a tracking function for users to view ongoing events that span several days. The tracking function presents related clusters over a three-day window. The system uses **NIA** to present an update summary containing only information that is fresh in the current day. Finally, DEMS produced the polished summaries.

The articles are processed as described in the earlier chapters. There are two notable differences:

1. The **NIA** system described earlier uses a learning algorithm to construct a classifier that identifies new passages. Now, this learned model is used as a classifier by the

system to identify the sentences that contain clauses with new information.

2. Instead of passing all the inputs to the learner, **NIA** gives the novel sentences to DEMS to give the user a summary of the desired size showing only the current developments.

The model is incorporated in the **NIA** code and only the sentences with novel material are output, and the summarizer runs independently. It is a sentence-extraction system, it is intended to be replaced in future work with a module that performs text generation.

9.1 DEMS

DEMS (Dissimilarity Engine for Multi-document Summarization) was built to cope with problems that new information poses for summarization. It uses several specially developed strategies to select interesting and informative sentences, including an innovative measure of importance derived from the analysis of a large news corpus. It also automatically selects which of several weight vectors to use in its sentence ranking by examining the cluster characteristics, such as a similarity matrix for the documents. The system also computes concept frequencies rather than word frequencies as an additional measure of importance. It merges these with a number of familiar summarization heuristics to rank sentences.

The summarizer currently produces a large portion of the regular summaries in Newsblaster. In addition, Columbia's submission to Document Understanding Conference since 2001 was mostly produced by DEMS. At the most recent DUC, in the summer of 2004, it performed quite well. In the evaluation by humans on content, it placed second on the 44 clusters it summarized in Task 2 general summaries[1] Table 9.1 shows the scores on these 44 clusters, which were lower than those on the remaining six, which were the clusters with the most similar documents.

[1]DEMS in normally paired with Multigen [MKH+99], which seeks to find the similarities across the documents in a cluster. Multigen is normally assigned the clusters in which the documents are most similar to one another.

Best System	0.29
DEMS	0.26
Average	0.2
Median	0.22

Table 9.1: DEMS performance at DUC 2004, as judged by the human assessor for overall coverage of model summaries written independently by humans.

9.1.1 Lead Values

The importance metric in DEMS is derived from the fact that the lead sentences of news articles – those in the first paragraphs of the articles – usually make excellent brief summaries [RBM94], but in multi-document summarization, the system has two or more, possibly many, lead sentences to choose from. The leads may be repetitive in content, but different on the surface; or there may be too many articles for all the lead sentences to fit in the summary. Further, some articles, particularly features, delay stating the point of the article until the fifth or sixth paragraph, and including such feature leads would put cryptic non sequiturs in the summary.

But, if the summarizer can identify paragraphs with information that is often found in the lead sentences, then such paragraphs are likely to contain important and interesting information. I examined first a large corpus of New York Times articles from 1996, and later a corpus of Reuters articles from the same year to determine what features could distinguish lead sentences from the average sentence [Sch02]. Using just the noninflected forms of the words as the features, I developed lists of 4,600 and 4,900 words from the two corpora that tended, with a reasonable measure of statistical significance, to be in the leads of articles more often than in the full text. I hypothesized that, on average, sentences with more high "lead words" would tend to reflect important events. Table 9.2 shows a sample of lead words from the Reuters corpus. The criteria for selecting the lead words is:

$$\frac{p(W_{inlead})}{p(W_{anywhere})} > 1$$

The ratios were checked for statistical significance with the binomial test and only those

Cynical	Eerie	Renovator
Conscription	Convalescent	Vial
Waterlogged	Showpiece	Extricate
Watershed	Rivet	Caravan

Table 9.2: A selection of lead words

with ratios where $pvalue < 0.05$ were accepted for inclusion in the lexicon.

The lead words are used as binary values, and averaged over the entire article, so that the sentence richest in lead words gets the highest score. By using the lexicon of lead words we are often able to locate secondary topics of interest in the articles, and to make comparisons of importance across documents. In a more general sense, the lead words allow passages with new information to be ranked higher, since the ranking does not depend on repetition within the cluster. I am, thus, making a distinction between *local importance*, that is importance in the context of the articles we have in a cluster, and *global importance*, or importance is some larger, universal context (See Figure 9.1).

9.1.2 Verb Specificity

In an effort to put sentences with the maximum amount of content into the summaries, I used the idea of verb specificity, developed in earlier work on a a biographical summarizer [SMC01]. In that work, I sought to retrieve from a large corpus a brief description and a short list of interesting events about a person or several people.

That system, Biogen, first extracted a short description of the person and used the head nouns in that description to select sentences with verbs closely associated with the kind of person the user was interested in, reasoning that these sentences would be more relevant.

In the biography work, I also experimented with the notion of verb specificity. If a verb was closely associated with only a few types of subjects, i.e. one that is highly specific, it would tend to convey information by itself in a sentence, and it would indicate a specific, well-defined event. For example, a verb like “arrest” suggests police activity. But less specific verbs (e.g., “be” or “do”) occur with a wide range of subjects and objects. The effort to find verbs that were indicative of the target individual was a precursor to my

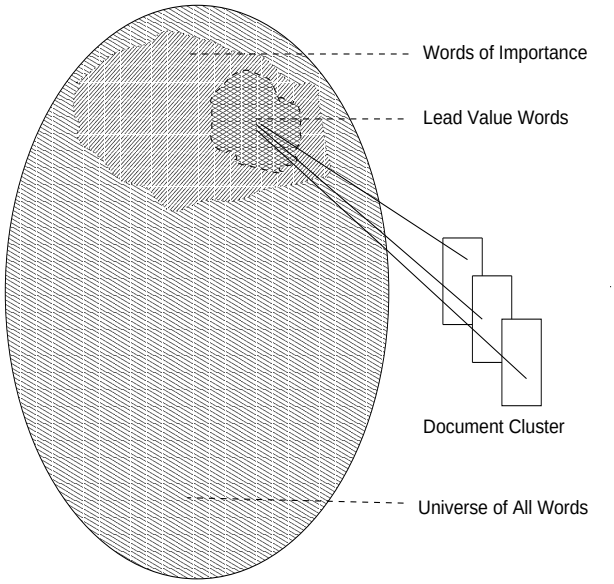

Figure 9.1: Consider words of importance to be a subset of all English words, and the lead words to be a subset of all importance words. Concentration of those lead words then should point to some of the interesting and important segments in the articles of a document cluster.

exploration of a promiscuity metric (See Chapter 5).

In constructing the biographical summarizer, descriptions could be found with the aid of pattern matching. But representative events important to that person were more difficult to select. In a large corpus, a great number of sentences might mention the particular name the user was interested in, but not all of them were interesting or informative. The system ranked sentences on the basis of how closely associated the verbs were to the terms in the person's description. The association of subject nouns to verbs was computed on the basis of a large corpus study. Mutual information statistics were collected from a year's worth of newswire.

The data suggested that many verbs were closely tied to only a few classes of nouns. The result was a "verb specificity" measure that reflects how often the mutual information between a particular verb and one noun or another exceeded a threshold.

$$VerbSpecificity = \frac{Count(V')}{Count(V)},$$

where V' is the occurrence of verb V with mutual information scores above a present threshold.

In the DEMS summarizer, the highest verb specificity in a sentence is used as the feature, in order to give increased weight to sentences rich in content. The motivation is to identify sentences that convey a complete thought by themselves, without depending too much on the surrounding context.

9.1.3 Cluster Classification

The summarization task at DUC motivated the strategy of categorizing different types of clusters and adjusting the DEMS parameters according to the cluster type. The first DUC, in 2001 revealed that there were three general types: 1.) single event tracked over a long period of time, usually about a particular person; 2.) multiple events of a similar nature; 3.) discussion of an issue with some related events. Examples of these three, respectively, are Elizabeth Taylor's bout with pneumonia, various marathon runners and races, gun control.

The categorization scheme isn't entirely clean, namely the sets are not entirely consistent. For example, many of the person-event (type 1) sets include one or two features about

the person, or the issue sets (type 3) may include one or two articles about a specific news event related to the topic.

It seemed clear that parameters should be tuned to the three types. Intuitively, type 1 requires some extra weight to both the main character (or entity) in the set and also needs to pay attention to the publication date so that the outcome is included in the summary. For example, did Elizabeth Taylor recover? Type 2 requires a broad brush approach, achieved by putting more emphasis on first sentences, and no emphasis on the target or publication date. Type 3 is more difficult, but experimentation showed that summaries were improved when the parameters emphasized the concepts most frequently found in the set. In all types of sets, it also seemed that summaries were more coherent if the outlier articles, those which didn't fit in the categorization scheme could be left out.

Thus the system creates a table of pairwise comparisons of the articles. Typical clustering techniques (those that are word-based and those that include all the content words) were too noisy. The table entries in DEMS is filled with equivalence class of words, and then only the n most frequent of them in each article. Using vector similarity over these n-element vectors provided a sharper indications of what each article in the set was about. The fit of each article to the set could be determined by combining the vector similarity scores along each row of the table.

When the span of similarity values is too wide, the set was usually type 2, multi-event. When the span was very narrow, it was either type 1 or type 3, and these could be distinguished easily by examining whether the most frequent concept was a named entity or not. For type 1 and type 3, I found that it was advantageous to drop the outlier stories – those that didn't quite fit into the set – from consideration. This helped make the summaries more coherent.

For type 1 summaries, the named entities and publication dates were important, but not to the extent we used them last year. For type 2 sets, the best sentences come from the beginning of the articles. In addition, outlier articles are ignored for type 1 and type 3 clusters. The categorization is made automatically, after the sentences are ranked but before the summaries are actually assembled.

9.2 Example

Figure 9.2 shows an example of typical NewsBlaster update summary applied to the news browsing system's tracking function. The summary is about 225 words in length, and was collected over three days at the end of March. The articles themselves were dated over five days, because many news sites link background articles to the current day's events. The cluster concerns the aftermath of the killing of Lebanon's Prime Minister Rafik Hariri on Feb 14, 2005. The articles are dated March 23, 24, 25, 26 and 27, and cover the release of an initial report on Hariri's assassination in February and two bombings at the end of March in Christian areas of Beirut.

Figure 9.3 shows the various events and the dates given on each report. Since Newsblaster conducts its web crawl late in the evening, articles reflecting the same information can have either the current date or a previous date. Most news sites do not time stamp their articles so that it is impossible to distinguish reports published early in the morning from those published late in the evening.

I extracted the rundown in Figure 9.3 manually, verifying that the full dates matched the days of the week given in each article. The summary reflects the information that is dated on March 26 and 27 in the time line. In general, each day's Newsblaster cluster on an important event spans two days, for sites before and after midnight, and often includes a number of background articles. The full texts of all the articles are in Appendix E. The summary contains mostly new information, although the presentation would clearly benefit from a better ordering approach. Of the eight sentences in the summary shown in Figure 9.2, five are clearly new, about the bombing on Saturday night – Sentences 4 through 8. Both clauses in Sentence 2 were judged to be new, and they contain details not found elsewhere in the article cluster, specifically that the three groups coexist, and that Syria's involvement could be characterized as a military occupation. Sentences 1 and 3 should not have been included. The bold type shows the clauses that were judged novel. In Sentence 1, the fault is the handling of numbers, which is left for future work, and in Sentence 3, the semantic lexicon was not able to handle the verb stymied. The program failed to equate that clause with earlier statements that the investigation was flawed and evidence was lost.

The model used to build this summary was adapted from the models trained on the

1. Since then, **hundreds of thousands of Lebanese have joined** demonstrations blaming Syria and demanding that it withdraw its 10,000 troops from Lebanon.
2.Lebanon's Christians, who coexist with large populations of Sunni and Shiite Muslims and Druse, are overwhelmingly opposed to Syria's military occupation.
3. This week, a United Nations investigation concluded that heavy-handed Syrian interference in Lebanese affairs had contributed to the polarized political atmosphere that led to Mr. Hariri's assassination, and that **a flawed local investigation had stymied efforts** to find the killers.
4. Fire engulfed two of the buildings, and the twisted wreckage of a car, presumably the carrier of the bomb, lay in the street.
5. A blast rocked Beirut on Saturday night, and Lebanese television broadcast stark images of severely damaged structures engulfed in flames.
6. Casualties cannot yet be confirmed in the blast, which is said to have occurred in eastern Beirut, a predominantly Christian area.
7. A print shop and possibly a timber yard were said to be affected, which would account for the extensive flames and smoke.
8. In Baghdad two months before, Iraqis in similar numbers had waited in line outside a high school to cast their ballots.

Figure 9.2: An example update summary from a recent NewsBlaster tracking graph. The summary contains about 225 words, as is standard in NewsBlaster. Two of the sentences were included although only part of them were judged new. The bold type shows the parts of these two mixed sentences that were judged new.

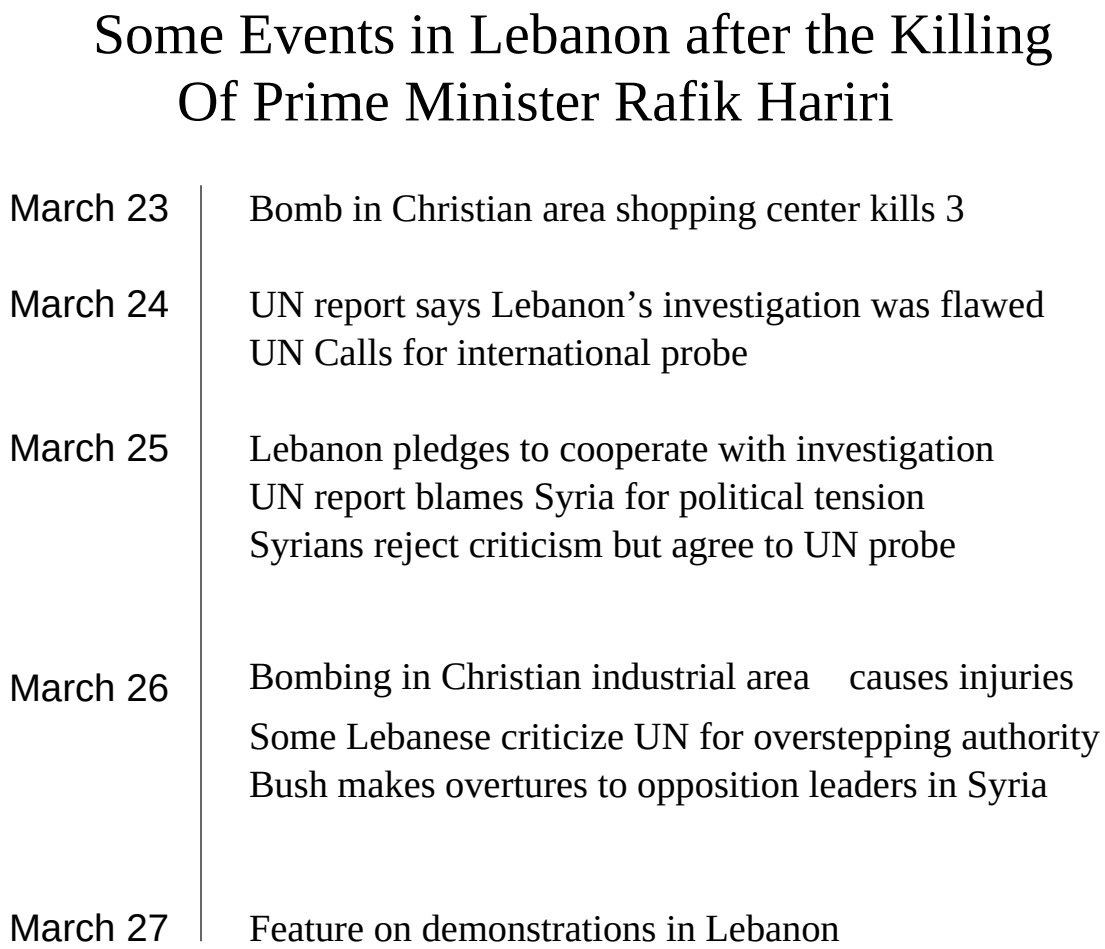

Figure 9.3: Major events reported by news web sites over three days of collection about the aftermath of the killing of the Lebanese Prime Minister Rafik Hariri.

TREC Novelty Track data (See Chapter 8. Like all the experiments in Chapter 8 the model was producer by the Weka [WF00] implementation of Ripper [Coh95]. The model could not be used directly because it used whole sentences for training, and thus many of the thresholds for attribute values were obviously too large for a clause-based analysis. Thus, the thresholds in the rule were experimentally adjusted. Using this model also allowed me to use plausible values for the docsim attribute, which were problematic, as explained in Chapter 8.

Chapter 10

Conclusion

Finding new-information in a collection of documents on a particular topic or event is a challenging and difficult problem, yet this thesis presents a method for highlighting just the new material with accuracy substantially above baseline approaches. The framework described above blends linguistic analyses with some innovative surface techniques to extract a wide array of features. Machine learning is used to combine these features into a classifier that partitions the input documents into novel and old segments. The novel segments are then fed into a summarizer and given to the user.

The system I have built, **NIA**, is a module to find novel information in a stream of documents. As texts are scanned, the information contained in them is added to the background, against which future documents will be filtered. **NIA** mainly uses clauses as units for its novelty decision, and identifies those which contain new facts. Some features extracted for each clause, other features look at the previous sentences. The system uses a standard implementation of rule induction, Ripper, to build the models later used to classify material.

NIA receives input from a clustering module, which is responsible for partitioning the inputs into coherent sets. It passes its output to a summarization module, which is responsible for shaping the output into a well-ordered, fluent document of the desired size. At present, **NIA** outputs all sentences that contain a novel clause, and passes these to the DEMS summarizer that I wrote.

In this research, I have shown that sentence by sentence comparisons are a poor choice for

new-information detection; indeed, at the TREC Novelty Track, a variety of pure sentence similarity metrics (considering novelty the inverse of similarity) have been used, but failed to achieve a precision much higher than the proportion of novelty in the test corpus. I created a small development and training corpus, and saw the same results in baseline tests, but the full system, with its learned models, can achieve substantially higher precision, without much loss of recall.

NIA itself is a pipeline of modules. The input cluster first undergoes syntactic and semantic analysis. From these, the two dozen features described in Chapter 7 are extracted. If the system is being run to build a classifier, the input needs to be enriched with gold-standard judgments for each clause unit, and the output, a list of feature vectors for each unit, are passed to a learner. If the system is being run to classify documents on line, the feature vectors are passed to the learned model and the associated passages are labeled as novel or old.

10.1 Contributions

The key contributions in this thesis are embodied in the development of **NIA**. A considerable amount of experimentation went into the development of the different stages of processing. The central notion underlying this research is that the addition of minimal syntactic, semantic and contextual knowledge will substantially surpass sentenced-based bag-of-words approaches on difficult problems such as the detection of new information. After extracting a set of more sophisticated features, I applied a variety of machine learning techniques to find the most appropriate for the task. In addition, the multi-document summarizer that **NIA** embodies several innovative techniques itself.

- *Clause-based features.* I developed a method of extracting features for new information from clauses. Once the sentences of the input texts are partitioned into separate clauses, pairs of words, which I call co-occurrences, are collected and used to determine novelty. These were described in Chapter 4 as the *Micro View*. If a pair has not been seen before, then it is evidence of novelty, otherwise it is evidence that the clause in question is not novel. The method is flexible in that the rules governing the pairings

are easily changed. The co-occurrences can be formed out of all the content words, without any further syntactic analysis once the clause boundaries have been found, or they can be restricted, for example, to only heads of phrases. The components of the co-occurrences can be weighted to reflect some notion of importance or topicality. I found that the simplest method, using all the words, worked best.

- *Use of context and discoure focus.* A key contribution of this work was the introduction of contextual information into the novelty classification. Chapter 6 discusses the genesis of the idea for the Text Retrieval Conference's Novelty Track, where the task was deciding whether whole sentences were novel or not. There, the system maintained a focus variable that was used to classify sentences that contained little evidence of being novel or not novel themselves. These were either sentences with few content words, or those with pronouns in prominent positions. Whenever a sentence showed strong evidence of novelty, the focus variable was updated. The notion of context was incorporated into the clause-based system by measuring the distance from the previous novel co-occurrence. The contextual features make up what I call the *Macro View*. These two views, or types, of features were combined into a coherent classifier with machine learning. After extensive experimentation, rule induction learners proved to be the most effective in the new-information task.

- *Combination of semantic resources.* I developed a way of combining semantic resources, whether manually created or automatically extracted. This addresses the problem of finding different realizations of the same underlying expression in the different input documents. With the database, all the words across all documents in the input set were put into referential equivalence classes. By assigning all common word tokens, nouns, verbs and adjectives into these equivalence classes, I am able to evaluate the novelty of co-occurrence beyond string matching. Experiments with various combinations of lexical resources showed that WordNet plus manually compiled information about nominalizations worked the best. Though the gains in performance were modest, it is, nonetheless, clear that more information leads to gains. Efforts to enlarge upon handbuilt resources were not too successful. I found that Dekang Lin's

existing lexicon of similar words proved to be too noisy to improve results. I chose his because it is publicly available and has been used by others. For proper nouns, I used a named-entity recognizer in a way to assign the same canonical names across all the documents.

- *Construction of a new-information corpus.* I built a corpus of pairs of news articles, in which novel passages, i.e. passages in the newer article that were not expressed in the older article. Each pair of articles was annotated by two people, journalism students at Columbia University. After the initial markup, the two people reviewed the differences between their decisions independently and were given the opportunity to revise their initial decisions. I found that this second look at the material eliminated most of their differences. The remainder of the differences were then negotiated by the two.

- *Exploration of a new metric for content.* I experimented with a new way to measure the information content of words in an effort to find a better way of determining the importance of the terms in a clause. A number of types of words are poor choices for the formation of co-occurrences: words that are very general, vague or serve some functional purpose. Because co-occurrences are restricted to the words contained in a single clause, it is important to avoid false alarms with words that are underspecified and do not convey content on their own. Stop words and the $TF * IDF$ metric from the information retrieval community are often used to discount the value of individual terms, but it was apparent that many low-content words are too infrequent to be identified by $TF * IDF$. I collected several statistics on the associations between pairs of words, and used machine learning to build a classifier to identity words like "treatment" or "type". I expected that eliminating co-occurrences with such "promiscuous words" would increase precision, and in many cases, this happened, but the corresponding drop in recall hurt overall results.

- *A new multi-document summarizer.* The DEMS multi-document summarizer was designed specificly with situations where new information should be treated prominently. Where many multi-document summarization approaches are guided by commonality,

DEMS tries to increase the ranking of infrequent statements provided that there is some evidence that they are nevertheless important or interesting. The summarizer contains a news-domain oriented feature that gives higher weights to words more often found at the beginning of articles, which is aimed at giving high rank to important words that have not been previously seen in the set. It is a sentence-extraction summarizer with several novel elements. The summarizer automatically assesses the cohesiveness of the set and chooses from among different ways of ranking the set.

10.2 Future Work

As this is a recently started area in Natural Language Processing, there is much to be done. With respect to **NIA**, a number of key points arise:

- One line of inquiry that should be addressed is the use of temporal information in the analysis. For one thing, events that occur after the date of previously processed events are clearly new. For another, different time stamps of similar events are good evidence of novelty. There is ongoing research on temporal annotation of text, and this should be incorporated in future work on new-information detection.

- Chapter 4 mentions making use of finer grained structural information, namely identifying the major constituents in a clause. I did some exploratory work on this, choosing the subject, object and verb of each clause, but the results fell short. The number of co-occurrences, the pairs of words that comprise the key features for classifying passages, were cut sharply, hurting recall. At the same time, the lack of complete noun phrase coreference diminished precision.

- One problem that is related to coreference issues is the appearance of ellipsis in text. Writers often splice clauses together without repeating the subject noun phrase. I implemented a surface method of identifying both the subject and object noun phrases and automatically attaching the most recent head to any succeeding subjectless clauses, but it failed to improve results. More experimentation is needed.

- The docsim feature, which measures the similarity of the current document to all

previous documents, was difficult to explore because of the small number of examples. I had approximately 2,400 clauses for training, but these came from only 31 articles. One solution would be to annotate more document pairs. Another would be to devise a way to use the sentence-based annotation of the TREC Novelty Track.

- Since the clause-based analysis did well in the new-information problem, it might improve the summarizer's choices of what to put in the final output.
- More work is necessary on weighting the words in the input documents, especially for building the co-occurrences. I tried different combinations of the promiscuous words feature described in Chapter 5 and on document frequencies from a large news corpus, but would like to explore this further and also to look at the idea of finding a set of core words for each set.

10.3 Limitations

NIA attacks a relatively high level task, and relies heavily on a number of systems to provide information about the input documents. Research on these other problems would fill several other theses themselves, and many of the limitations involve the limitation of available software used to analyze the input documents.

Language Resources

- Co-reference is a problem. Despite a large body of work on pronoun resolution, I showed in Chapter 4 that available software adds too much noise to justify its use. In addition, there are many other types of reference. While there are good named-entity recognizers, the generation that is currently available requires some string matching, but there are some government initiatives seeking improvements in this area.
- In semantics, WordNet is an invaluable resource, but it has numerous problems that were discussed in Chapter 5. The idea of automatically extracting semantic relationships from large corpora is an active area of research, but currently available resources add too much noise.

- Resources to deal with deeper semantic relationships – like semantic roles and disambiguation – as well as pragmatics and world knowledge are not available.

User Choices For this thesis, I paid no attention to user characteristics. Different users might have different emphases both in terms of the novelty classification and in the final selection of the summaries.

Domain Independence The system design is not directly domain dependent, but many of the resources used in the early stages of processing were trained in the news domain. These include the parser and the named-entity recognizer. Some domains, especially those that use a sublanguage and those that are informal might require new tools.

Text Generation The transformation of the **NIA** output to fluent text would be highly desirable but is beyond the scope of this thesis. New information is a difficult case for generation methods that rely on similarity across documents, since by definition, the system is looking for material that is mentioned for the first time.

Input Clustering **NIA** is dependent on the quality of the clustering algorithm that prepares the input. While automatic methods are well developed, errors can make new-information detection impossible in some cases.

Bibliography

[AAMH01] E. Agirre, O. Ansa, D. Martinez, and E. Hovy. Enriching wornet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other lexical Resources*, 2001.

[Abn96] Steven Abney. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*, 1996.

[AJAC+04] Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Donald Metzler, Mark D. Smucker, Trevor Strohman, Howard Turtle, and Courtney Wade. Umass at trec 2004: Notebook. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

[AKA91] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[And04] Rie Kubota Ando. Semantic lexicon construction: Learning from unlabeled data via spectral analysis. In *Proceedings of CoNLL-2004*, 2004.

[AWB03] James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the ACM SIGIR conference on research and development in information retrieval*, 2003.

[BBC+04] Stephen Blott, Oisin Boydell, Fabrice Camous, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Noel Murphy, Noel O'Connor, Alan F. Smeaton, Barry Smyth, and Peter Wilkins. Experiments in terabyte searching, genomic

retrieval and novelty detection for trec-2004. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

[BFL98] Colling F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 1st Annual Meeting of the COLING-ACL*, 1998.

[BL03] Regina Barzilay and Lillian Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of Human Language Technologies Conference*, 2003.

[BM01] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings European Association for Computational Linguistics 2001*, 2001.

[Bre96] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Bur98] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[CDO03] John M Conroy, Daniel M. Dunlavy, and Dianne P. O'Leary. From trec to duc to trec again. In *TREC Notebook Proceedings*, 2003.

[CEL95] CELEX. *The CELEX lexical database — Dutch, English, German.* Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen, 1995.

[Cha00] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the NAACL-2000*, 2000.

[CM02] J.R. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, 2002.

[Coh95] William W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[Col96] Michael Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the ACL*, 1996.

[Con04] John M. Conroy. A hidden markov model for trec's novelty task. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

[Cra99] Mark Craven. Learning to extract relations from medline. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[CT95] John G. Cleary and Leonard E. Trigg. K*: an instance-based learner using an entropic distance measure. In *Proc. 12th International Conference on Machine Learning*, 1995.

[CTOZC02] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection and named-page finding. In *Proceedings of the 11th Text Retrieval Conference*, 2002.

[DDJ+03] D.M.Dunlavy, D.P.O'Leary, J.M.Conroy, J.D.Schlesinger, S.A.Goodman, and M.E.Okurowski. Performance of a three-stage system for multi-document summarization. In *Proceedings of the Workshop on Text Summarization*, 2003.

[Dom99] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, 1999.

[DP97] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[DRRB00] Hongyan Jing Dragomir R. Radev and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extaction, utility-based evaluation and user studies. In *Proceedings of ANLP/NAACL-2000 Workshop on Automatic Summarization*, 2000.

[Dun93] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[ESL+03] David Eichmann, Padmini Srinivasan, Marc Light, Hudong Wang, Xin Ying Qiu, Robert J. Arens, and Aditya Sehgal. Experiments in novelty, genes and questions at the university of iowa. In *TREC Notebook Proceedings*, 2003.

[EZB+04] David Eichmann, Yi Zhang, Shannon Bradshaw, Xin Ying Qiu, Li Zhou, Padmini Srinivasan, Aditya Kumar Sehgal, and Hudon Wong. Novelty, question answering and genomics: The university of iowa response. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

[Fre95] Yoav Freund. Boosting a weak learning algorithm by majority. In *Proceedings of the Workshop on Computational Learning Theory*, 1995.

[FS96] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 1996.

[FSZC99] Wei Fan, S.J. Stolfo, J. Zhang, and P.K. Chan. Adacost: Misclassification cost-sensitive boosting. In *Proceedings of the International Conference on Machine Learning*, 1999.

[GC98] Jade Goldstein and Jaime Carbonell. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, 1998.

[GCY92] W.A. Gale, K.W. Church, and D. Yarowsky. One sense per discourse. In *Proceedings of the DARPA speech and natural language workshop*, 1992.

[GJ02] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[GMCK00] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL-2000 Workshop on Automatic Summarization*, 2000.

[GPDR01] Rebecca Green, Lisa Pearl, Bonnie J. Dorr, and Philip Resnik. Lexical resource integration across the syntax-semantics interface. In *Proceedings of the NAACL workshop on WordNet and Other lexical Resources*, 2001.

[Gre94] Gregory Grefensette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.

[GS86] Barbara J. Grosz and Candace L. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[HKH+01] Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. Simfinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*, 2001.

[HM93] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, 1993.

[HMM99] Sanda Harabagiu, George Miller, and Dan Moldovan. Wordnet 2 - a morphologically and semanically enhanced resource. In *Proceedings of SIGLEX-99*, 1999.

[HMM+01] S. Harabagiu, D. Moldovan, P. Morarescu, F. Lacatusu, R. Mihalcea, V. Rus, and R. Girju. Gistexter: A system for summarizing text documents. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[HT03] H. Van Halteren and S. Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT/NAACL-2003 Workshop on Automatic Summarization*, 2003.

[IH01] Diana Zaiu Inkpen and Graeme Hirst. Building a lexical knowledge-base of near-synonym differences. In *Proceedings of the NAACL workshop on WordNet and Other lexical Resources*, 2001.

[ILK+03] Abraham Ittycheriah, Lucian Lita, Nanda Kambhatla, Nicolas Nicolov, Salim Roukos, and Margo Stys. Identifying and tracking entity mentions in a maximum entropy framework. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003.

[JJDM01] J.M.Conroy, J.D.Schlesinger, D.P.O'Leary, and M.E.Okurowski. Using HMM and logistic regression to generate extract summaries for duc. In *Proceedings of the Workshop on Text Summarization*, 2001.

[JJJD04] J.M.Conroy, J.D.Schlesinger, J.Goldstein, and D.P.O'Leary. Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Workshop*, 2004.

[JL95] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[JM00] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice-Hall, 2000.

[Joa98] Thorsten Joachims. Making large-scale svm learning practical. Technical Report LS8-Report 24, Universitat Dortmund, 1998.

[Kle02] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[KPM02] Paul Kingsbury, Martha Palmer, and Mitch Marcus. Adding semantic annotation to the Penn TreeBank. In *Proceedings of Human Language Technologies Conference*, 2002.

[KSC+03] Srikanth Kallurkar, Yongmei Shi, R. Scott Cost, Charles Nicholas, Akshay Java, Christopher James, Sowjanya Rajavaram, Vishal Shanbhag, Sachin

Bhatkar, and Drew Ogle. Umbc at trec 12. In *TREC Notebook Proceedings*, 2003.

[Lew98] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15, 1998.

[LH00] Chin-Yew Lin and E.H. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics*, 2000.

[LH01] C-Y Lin and E. Hovy. Neats: A multidocument summarizer. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[Lin98a] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, 1998.

[Lin98b] Dekang Lin. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, 1998.

[Lin98c] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, 1998.

[Lin04] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, 2004.

[LJ80] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[Mar01] D. Marcu. Discourse-based summarization in duc-2001. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[MB97] Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings, American Association for Artificial Intelligence 1997*, 1997.

[MB99] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67, 1999.

[MBC+03] Kathleen McKeown, Regina Barzilay, John Chen, David Elson, David Evans, Judith Klavans, Ani Nenkova, Barry Schiffman, and Sergey Sigelman. Columbia's NewsBlaster: New features and future directions. In *Proceedings of Human Language Technologies Conference*, 2003.

[MBE+02] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's NewsBlaster. In *Proceedings of Human Language Technologies Conference*, 2002.

[MBF+90] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312, 1990.

[MGM+98] Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth reeves. Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX'98*, 1998.

[MKH+99] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. Towards multidocument summarization by reformation: Progress and prospects. In *Proceedings of American Association for Artificial Intelligence 1999*, 1999.

[MM01] Rada Mihalcea and Dan I. Moldovan. extended wordnet: Progress report. In *Proceedings of the NAACL workshop on WordNet and Other lexical Resources*, 2001.

[Mor83] William Morris. *The American Heritage Dictionary.* Houghton Mifflin Co., 1983.

[Mor00] Thomas S. Morton. Coreference for NLP applications. In *Proceedings of the 38th Annual Meeting of the ACL*, 2000.

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn TreeBank. *Computational Linguistics Special Issue on Using Large Corpora*, 19(2):313–330, 1993.

[NP04] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: the pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2004.

[Pla99] J.C. Platt. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, pages 185–208. MIT Press, 1999. In B. Scholkopf and C. Burges and A. Smola, editors, *A*dvances in Kernel Methods – Support Vector Learning.

[PTL93] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the ACL*, 1993.

[Qui93] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[RBM94] L.F. Rau, R. Brandow, and K. Mitze. Domain-independent summarization of news. In *Summarizing Text for Intelligent Communication*, 1994.

[Res99] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[Ril96] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Artificial Intelligence*, 1996.

[Ris01] Irina Rish. An empirical study of the naive bayes classifier. In *Proceedings of the workshop on Empirical Methods in AI at IJCAI-01*, 2001.

[RJB00] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Summarization of multiple documents: clustering, sentence extraction. In *Proceedings, ANLP-NAACL Workshop on Automatic Summarization*, 2000.

[R.L90] John R.L.Bernard. *The Macquarie Encyclopedic Thesaurus*. The Macquarie Library, 1990.

[Rog11] Peter Roget. *Thesaurus of English Words and Phrases*. Longmans, Green and Company, 1911.

[RWC97] Yael Ravin, Nina Wacholder, and Misook Choi. Disambiguation of proper names in text. In *Proceedings of the 17th Annual ACM-SIGIR Conference*, 1997.

[RZZM04] Liyun Ru, Le Zhao, Min Zhang, and Shaoping Ma. Improved feature selection and redundancy computing – thuir at trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

[Sch02] Barry Schiffman. Building a resource for evaluating the importance of sentences. In *Proceedings of the Language Resources and Evaluation Conference*, 2002.

[Sid03] Advaith Siddharthan. Resolving pronouns robustly: Plumbing the depths of shallowness. In *Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2003.

[SJM+02] J.D. Schlesinger, J.M.Conroy, M.E.Okurowski, H.T.Wilson, D.P.O'Leary, A.Taylor, and J.Hobbs. Understanding machine performance in the context of human performance for multi-document summarization. In *Proceedings of the Workshop on Text Summarization*, 2002.

[SM00] Barry Schiffman and Kathleen McKeown. Experiments in automated lexicon building. In *Proceddings of COLING-00*, 2000.

[SMC01] Barry Schiffman, Inderjeet Mani, and Kristian J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings European Association for Computational Linguistics 2001*, 2001.

[SMR99] Kristi Seymore, Andrew McCallum, and Ronald Rosenfeld. Learning hidden markov model structure for information extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[SNM02] Barry Schiffman, Ani Nenkova, and Kathleen McKeown. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*, 2002.

[Sob04] Ian Soboroff. Draft overview of the trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

[SpZ+03] Jian Sun, Wenfeng pan, Huaping Zhang, Zhe Yang, Bin Wang, Gang Zhang, and Xueqi Cheng. Trec-2003 novelty and web track at ict. In *TREC Notebook Proceedings*, 2003.

[Str18] William Strunk. *The Elements of Style*. Bartleby.com, 1918. Online edition: http://www.bartleby.com/141/.

[Tve77] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[Uni03] Meiji University. Meiji university web and novelty track experiments at trec 2003. In *TREC Notebook Proceedings*, 2003.

[Vap95] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[War96] Grady Ward. Moby thesaurus, 1996. Moby Project.

[WCN+01] Michael White, Claire Cardie, Vincent Ng, Krii Wagstaff, and Daryl McCullough. Detecting discrepancies and improving intelligibility: Two preliminary evaluations of riptides. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[WF00] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 2000.

[WM95] David Wolpert and William G. Macready. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

[WM97] David Wolpert and William G. Macready. No free lunch theorems for optimization. In *IEEE Transactions on Evolutionary Computation*, 1997.

[YW03] Y. Yang and G.I. Webb. On why discretization works for naive-bayes classifiers. In *Proceedings of the 16th Australian Conference on AI*, 2003.

[YZCJ02] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. Topic-conditioned novelty detection. In *Proceedings of the Special Interest Group in Knowledge Discovery and Data Mining*, 2002.

[ZMDP95] Wendy M. Zickus, Kathleen F. McCoy, Patrick W. Demasco, and Christopher A. Pennington. A lexical database for intelligent AAC systems. In *Proceedings of RESNA '95 18th Annual Conference*, 1995.

[ZXB+04] Hua-Ping Zhang, Hong-Bo Xu, Shuo Bai, Bin Wang, and Xue-Qi Cheng. Experiments in the trec 2004 novelty track at CAS-ICT. In *The Thirteenth Text Retrieval Conference (TREC 2004) Notebook*, 2004.

Appendix A

Model Summaries on the Euro

These are the four model summaries, written by the NIST assessors, for the DUC cluster D30033 at the 2004 evaluation. The 'M' indicates that the summaries are multi-document summaries, taken from a cluster of 10 news articles, and the 100 is the maximum number of words allowed. The last letter is a code for the different assessors. The sentences in the summaries are split into "model units", which are often the clauses that **NIA** uses.

A.1 D30033.M.100.T.A

Eleven countries were to adopt a common European currency , the euro , on Dec. 31 , 1998 .

In November and December there were various reactions .

France made moves toward a pan-European equity market .

Ten of the countries quickly cut interest rates

causing fear of overheating in some economies .

In Denmark , which had earlier rejected the euro , a majority was now in favor .

And in faraway China , the euro was permitted in financial exchanges .

Whatever the outcome , the euro 's birthday , Dec. 31 , 1998 , would be an historical date .

Some saw it as a step towards political union

while others already considered themselves as citizens of Europe

A.2 D30033.M.100.T.D

Eleven European nations are forming a " Euro zone " .
Britain , Denmark , Sweden , and Greece are not part of it .
Danes favor joining .
The Euro became official for intergovernmental transfers on Dec 31 , 1998 ,
but bills and coins will not come until 2002 .
The Paris , London , and Frankfurt stock exchanges have formed an alliance .
Euro nations cut interest rates
and inflation fell to an average 0.9
China has authorized use of the Euro in trade .
The president of the European Central Bank warns
that growth is slowing
and that he plans to complete his term .
The EU monetary action has given rise to the new mobile , multi-lingual , non-na

A.3 D30033.M.100.T.E

France 's offer
to host a financial meeting for nine other European nations is seen as a precursor to a pan-European market .
It shows how the new currency , the euro , is reshaping Europe financially .
Eleven European nations lowered key interest rates in preparation for the conversion .
China made trading in euro official Monday
when it accepted its use in trade and finance starting Jan. 1 .
Denmark and Sweden may not join the euro for political reasons .
Some smaller nations may become unstable from a growing inflation decline .
A new generation , already cosmopolitan , will n't be shocked .
The head of the new European Central Bank will not s

A.4 D30033.M.100.T.G

On 1 Jan 1999 , the euro , a currency serving 11 European nations , entered the world financial market .
As time grew short , questions remained over the pan-European market .

When the head – serving an 8-year term – of the European Central Bank ,
which governs the euro ,
expressed fear of a slowing economy ,
the nations simultaneously dropped interest rates ,
spurring the market .
Annual inflation rates also were encouraging .
Denmark , who along with Sweden and Britain eschewed the euro ,
was becoming interested .
As a step to a unified Europe , the euro will well serve the new , mobile , multi-lingual , business generation and could po

Appendix B

Annotators' Instructions

My development corpus consisted of 31 pairs of documents, with each pair on a particular news event. Each pair was shown to two annotators – all six annotators were students at the Columbia School of Journalism. Figure B.1 is the web page that explained what the background was, and Figure B.2 what the new article was and how to mark it up.

The annotation was done over the web, requiring the students simply to highlight a section of text and choose categories of novelty and importance. After their markup, there was a two-stage negotiation process. First, each of annotator was shown the differences and asked to reconsider the passages in contention. If any disagreement was left, they were then asked to discuss their differences by email and resolve them.

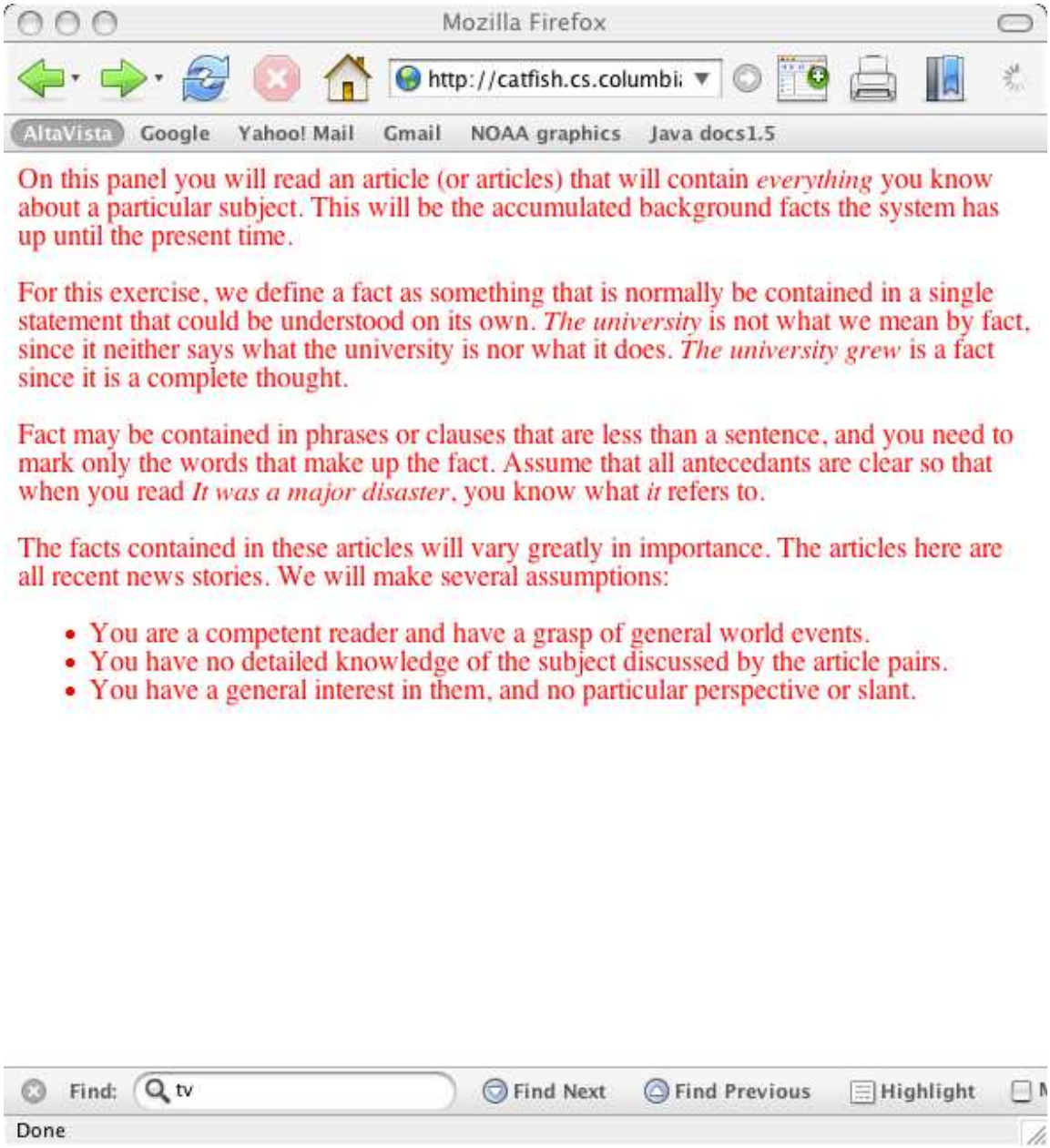

Figure B.1: Explanation of background for annotators

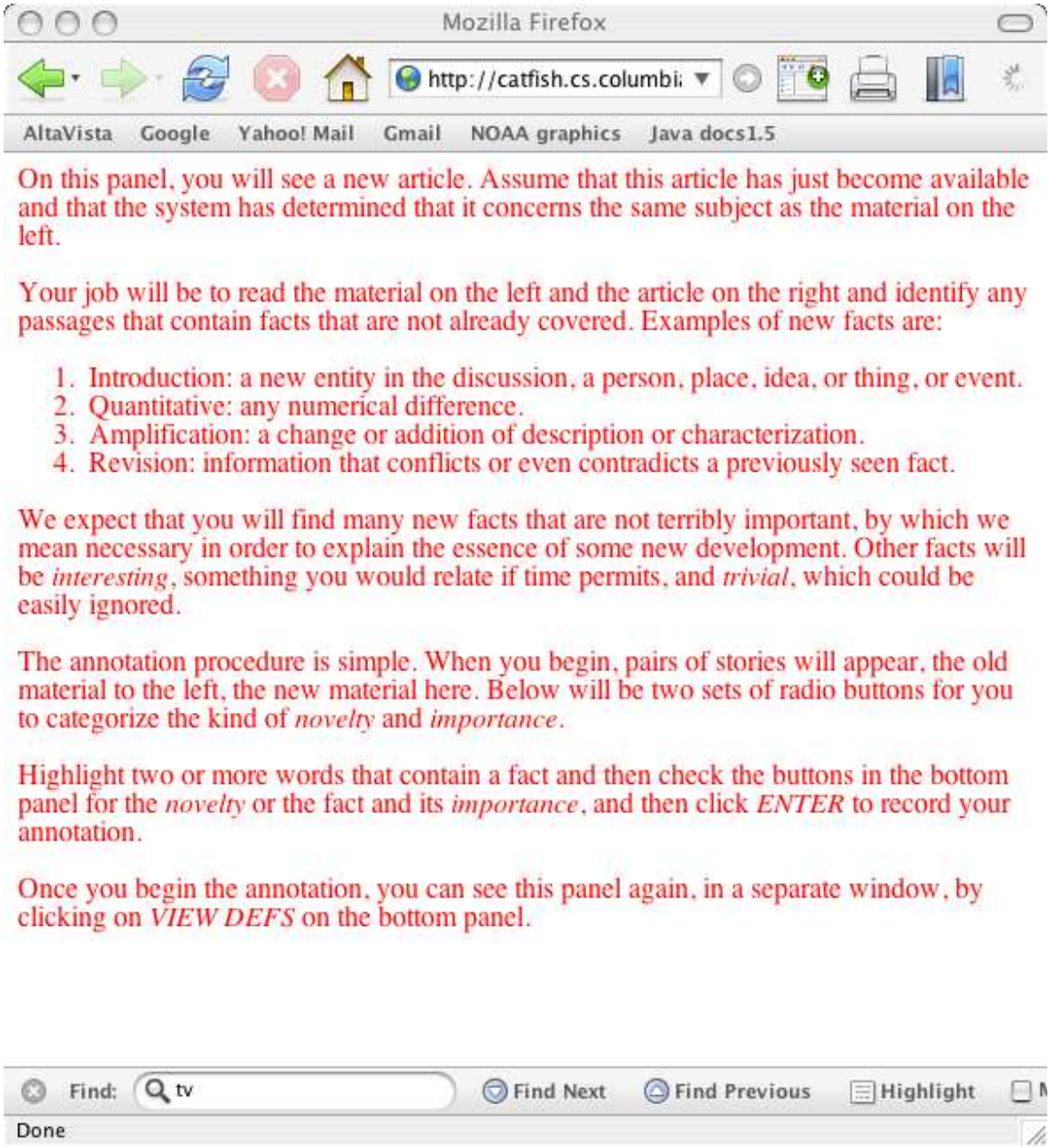

Figure B.2: Instructions on markup for annotators

Appendix C

Training Corpus

Here are the 31 pairs of articles used for system development and training. For each pair, first is the background article, and then the new article. The segments the annotators agreed were novel appear in bold in the new articles.

C.1 amnesty background

Geneva - China executed more people than any other country last year - two-thirds of the known world total of more than 1,500 - and many of those cases violated international law, the human rights group Amnesty International said Friday.

The group also said the United States was the only country that executed offenders who were under 18 when they committed their crimes. Three such offenders were executed in Texas last year.

"This blight on our country's human rights record belies our claim to be an international human rights defender," said William F. Schulz, executive director of Amnesty International USA.

Amnesty International released its annual report in Geneva to coincide with the meeting of the 53-nation Human Rights Commission - the top United Nations human rights body. Amnesty International opposes the death penalty in all cases.

At least 1,526 people were executed in 31 countries last year, the group said. China accounted for at least 1,060 of those, followed by Iran with 113, London-based Amnesty International said. However, the true number in both countries was believed to be much higher. "Many cases were in blatant violation of international standards on the application of the death penalty," Amnesty International said. "Prisoners were sentenced to death following unfair trials."

The United States had the third-highest number of executions.

Ray Krone, who spent more than 10 years in an Arizona prison - nearly three of them on death row - for the murder of a barmaid before DNA evidence cleared him, told reporters many innocent people in the United States are convicted and executed because they do not have the money for proper legal representation.

"I was naive, ignorant of how the system really worked. I believed innocence was protection, I thought the truth would be forthcoming. I am proud of our country, but this is something that has to change," said Krone, who was in Geneva to speak against the death penalty before the commission.

Amnesty said there has been progress toward abolition, with 111 countries having abolished the

death penalty"in law or practice" as of the end of 2002.

C.2 amnesty new article

THE NUMBER of people executed worldwide halved last year– but the United States bucked the trend and put to death more prisoners .

Amnesty International reported that at least 1,526 people were executed in 31 countries in 2002 **, compared with more than 3,000 the year before**.

By far the highest number of death penalties were recorded in China (1,060) , followed by Iran (**113 , including a man and a woman reportedly stoned to death**) , and the United States (**71**) **.**

These three countries accounted for more than 80 per cent of known executions in 2002.

The American toll , which including three juvenile offenders **, was higher than in 2001 , when 66 prisoners were executed**.

Releasing its figures , the human rights organisation called for a moratorium on capital punishment **.**

Confirmed numbers of prisoners put to death were also recorded in Saudi Arabia (48) , Sudan (40) , Vietnam (34) , Tajikistan (28) , Egypt (17) and Jordan (14) .

In each of these countries the actual figure could be far higher .

Amnesty also said it believed there had been " scores " of judicial executions in Iraq .

Despite the international fall in numbers of those executed last year , Amnesty said they should be seen as minimum figures **because of the secrecy over the death penalty in many countries**.

It said the number of judicial executions was higher than in 2,000 **.**

Amnesty said it also recorded at least 3,248 instances in 67 countries of people being sentenced to death .

They included Amina Lawal, a Nigerian mother-of-four , who still faces stoning for adultery , and 88 people in Sudan , including two 14-year-old boys **.**

Kate Allen , Amnesty UK director , said : " Last year figures show a significant drop in the use of this cruel and unnecessary punishment but worryingly , executions are higher now than they were at the start of the millennium . "

She singled out the US for criticism of its " shameful record " of " sub-standard trials " leading to death sentences for juvenile offenders and the mentally ill .

Amnesty said **111 countries had scrapped the death penalty**in law or practice .

In 2002 **, Cyprus and the Federal Republic of Yugoslavia (now Serbia and Montenegro) abolished it for all offences**

C.3 aolspam background

American Online fought back yesterday against junk E-mail, suing five marketers for swamping its subscribers with 1 billion messages hawking everything from pornography to steroids.

The suits, filed in federal court in Virginia, seek $10 million in damages and court orders blocking the defendants from sending any more of the E-mails, which are known as spam.

AOL and other Internet service providers have been waging an uphill battle against spam, which has spread like some kind of electronic kudzu. Nearly 41% of all E-mails sent in the U.S is spam, a 500% increase in the last 18 months.

The E-mails are a growing nuisance to Internet users, and the defendants named by AOL allegedly sent the kind of obnoxious E-mails known to anyone who's cleared an in-box.

One defendant, George Moore, of Maryland Internet Marketing, hawked mortgages, computer software plus a host of other products and services.

"Protect yourself against harmful viruses... This is a MUST for ALL computer users!!!" one of his E-mails began.

"Build huge muscle and rapidly burn body fat today," said another E-mail cited in the suit, promoting steroids from Mexico.

Defendant, Michael Levesque of Washington state allegedly sent spam advertising lotsofnudeslutz.net by touting it as the address to visit"if you are tired of boring so-so porn sites."

Some spammers employed many of the fraudulent methods used by junk E-mailers, including placing bogus return addresses on their messages.

But Moore has a real address in Linthicum, Md. Online he goes by the E-mail name of Dr. Fatburn and has used E-mailing to promote such products as Extreme Colon Cleanser, FAT-N-EMY and Extreme Power Plus.

"It's not constructive to give any kind of comment," he told the Daily News. The man accused of dumping millions of unwanted E-mails into computers added, "I will respect you not calling us in the future."

Three of the defendants are named as"John Doe," meaning AOL could not determine their identities. By filing the lawsuits, AOL will now have the authority to issue subpoenas to try to track them down.

The unsolicited E-mails have become costly to companies and Internet providers like AOL and Earthlink, which are spending billions of dollars a year in an uphill battle to block them.

Since the mid 1990s, AOL has filed 100 lawsuits seeking to stop junk E-mail, spokesman Nicholas Graham said. Yesterday's suits were marked by the massive scope of the spamming.

The suits also are the first to cite complaints lodged by AOL members who use the"report spam" feature in AOL 8.0. The defendants allegedly sent spam that resulted in 8 million such complaints by AOL users.

The new suits also were launched with unusual publicity, to gain the attention of AOL subscribers as well as spammers, Graham acknowledged.

After years of growth, AOL has been losing subscribers and is under pressure from its remaining customers to stop the spam scourge.

The suits allege the spammers violated the Computer Crimes Act in Virginia; the Federal Computer Fraud and Abuse Act, and a law in Washington state.

Computer Software: "Protect yourself against harmful viruses... This is a MUST for ALL computer users!!!"

Other Products and Services:

Extreme Colon Cleanser

C.4 aolspam new article

America Online Inc. said yesterday that it has filed five lawsuits - including one against a Linthicum man and his company - alleging that businesses and individuals have illegally barraged customers with junk e-mail , known in the industry as spam **.**

The Dulles , Va.-based Internet companyalleges that the defendants have sent about a billion spam e-mail messages that include pornography , offers to mortgage or refinance homes , and offers to buy steroids , college degrees or software .

It also alleges that the defendants sent the e-mail illegally , for instance , by dodging spam filters or falsifying e-mail addresses .

" In the case of these guys , we already told them that they are not authorized to send mail on our network , and they continue to do it **," said Randall Boe , general counsel for America Online.**

The lawsuits were filed in U.S. District Court for the Eastern District of Virginia in Alexandria because the company e-mail network is based in Virginia.

The five complaints against 12 defendants are the **latest in several rounds of complaints**AOL has filed against spammers since 1998 .

Boe said the company has sued more than 100 spammers **and never lost a case**.

The company has barred people from sending unsolicited e-mail and collecting tens of millions of dollars , Boe said .

In the lawsuits , AOL asked for at least $10 million in damages and court orders for the defendants to stop sending spam .

Three of the lawsuits were filed against " John Does . "

Boe said that AOL expects to identify them soon .

George Moore of Maryland Internet Marketing Inc. of Linthicum was named as a defendant .

" Our complaint against him and his company is that he has run a program where he and affiliates of his are sending millions of pieces of e-mail promoting things like anti-virus software , mortgage leads ," Boe said .

" That generated a couple of hundred thousand complaints from AOL members .

We 've asked him to stop , and he did **n't . "**

Moore could not be reached for comment yesterday.

Another defendant is Michael Levesque and his firm , Byte Night LLC , a Seattle company .

AOL lawyers believe Levesque is affiliated with Moore .

Levesque attorney , Derek Newman of the Seattle law firm Newman and Newman LLP , said that Byte Night is a holding company for some of Levesque investments and his client did not send out spam.

" The claims against him are without merit , and we 'll contact AOL and hopefully they'll dismiss ," Newman said , " and if they do n't dismiss it , we 'll defend it ."

Boe said that America Online chose the defendants based on information from its members .

When AOL customers hit a " report spam " button their screen , the piece of junk e-mail is sent to an America Online database that is used for investigations .

AOL said the defendants ' spam prompted more than 8 million complaints from its members .

" We erect filters that are designed to block junk e-mail , so what (spammers) do is either hack their way in to using other servers so the mail appears to be coming from a place that completely unrelated to their business , or they will falsify parts of the e-mail ," Boe said

---

C.5 arnold background

Buoyed by his high-profile backing of the victorious Proposition 49, Arnold Schwarzenegger appears to be the best hope of the Grand Old Party to wrest the governor's seat from Democrats four years from now.

With more than $3 million in commercials featuring his mug and touting the mom-and-apple-pie proposition to boost after-school programs for kids, the actor from Austria helped transfer his screen credibility to politics.

"Arnold was really the only winner on the Republican side," said Mark Baldassare, pollster for the Public Policy Institute of California. "His proposition got 3.5 million votes. It's hard to find another Republican in California that was as successful."

For months, veteran Democrats Attorney General Bill Lockyer and Treasurer Phil Angelides have been voraciously raising campaign cash, even though they faced token GOP opposition this year.

Both were trying to amass as big a war chest as they could for a gubernatorial run before tight contribution limits took effect Wednesday.

By election day, Lockyer had banked about $10 million.

"I was a Boy Scout; you're supposed to be prepared," Lockyer said.

Angelides had $7.4 million as of Oct. 19, but late contribution reports show him dumping $1 million of his own money into his campaign account, an action prohibited under the new limits.

Lockyer was the biggest vote-getter on the Democratic ticket, the only one to crest 50 percent and capturing 150,000 more votes than Gov. Gray Davis.

Right behind Lockyer was Lt. Gov. Cruz Bustamante who at more than $3.2 million got 100,000 more votes than Davis.

But Bustamante, a not-very-effective fund raiser, spent most of his campaign kitty to stave off a challenge from GOP Sen. Bruce McPherson.

Newly elected Insurance Commissioner John Garamendi has run for governor twice – the second time after one term as commissioner in 1994.

Steve Westly, the multimillionaire Democrat narrowly elected controller, could also be a candidate.

None of the candidates speaks publicly of gubernatorial aspirations.

"The best politics is good government," Lockyer said. "I'm going to be the best possible attorney general I can, and if that leads to promotional opportunities with the voters, they'll let me know."

Angelides says he just wants to focus on his current job but does pick a high-profile priority.

"My No. 1 priority is to keep being an active leader on the national scene in pushing forward corporate reform, particularly in light of the Bush ascendancy," Angelides said.

A woman could jump into the gubernatorial race and take the Democratic nomination.

"In a crowded field, a substantive woman candidate would stand a good chance, and it didn't go unnoticed there weren't any women on the Democratic ticket this time," said Gale Kaufman, a Sacramento political consultant.

With all GOP candidates for statewide office being electorally wiped out, the state's ranking Republican is Senate Minority Leader Jim Brulte of Rancho Cucamonga. Holding that position ensures his name will be mentioned as a gubernatorial possibility.

And Bill Simon, the GOP challenger to Davis this year, can't be ruled out. Simon has hinted that he wants to stay in politics, and his surprisingly good showing in Tuesday's election keeps him in the mix.

Schwarzenegger is not without sound political guidance.

The team behind Prop. 49 – and Schwarzenegger – includes old GOP hands like Bob White, former Gov. Pete Wilson's chief of staff, and George Gorton, Wilson's chief political consultant.

But at least for now Schwarzenegger benefits mainly from his celebrity – and that helps his short-term game plan.

"People feel they know him because of his movies, and that helps in transferring his credibility to a new environment," said Barbara O'Conner, a political communications professor at Cal State Sacramento. E-mail Greg Lucas at EMAIL

C.6 arnold new article

Arnold Schwarzenegger came to Chapman University shortly before the election to discuss his after-school initiative for MSNBC " Hardball " talk show .

Schwarzenegger told " Hardball " host Chris Matthews that the last time he visited Chapman – pick up an honorary degree– " I said when I left , ' I 'll be back ' Right , well , I 'mback . "

He 'll be back again .

Though he had to cancel this month **appearance at the Orange County Performing Arts Center**, people in the area will see a good deal of him in the years ahead .

And not just in the theaters .

On election night **, Schwarzenegger was a bigger winner than Gray Davis**.

Not only did his ballot proposition triumph , but the GOP poor showing left the party longing for a champion .

In four years , Republicans may turn to him as their candidate for governor .

If they do , Orange County will be one of his major bases of support **.**

Some local conservatives grumble about his support for abortion rights , but he is immune to their favorite epithet for Republican moderates .

Nobody calls this man a "squish . "

Schwarzenegger has qualities that the state party has been missing .

The first is a reputation for " compassionateconservatism . "

He has **long worked on children causes , such as the Special Olympics and the Inner-City Games .**

The first President Bush named him to head the President Council on Physical Fitness and Sports , where he became a national spokesman for physical education .

This record made him a credible sponsor of the initiative .

Significantly , the proposition Web address was www.joinarnold.com .

After-school programs **are a popular idea**, so the measure drew a rainbow coalition of endorsements .

Orange County supporters **included Republican Rep. Dana Rohrabacher , state Sen. Dick Ackerman , and Sheriff Michael S. Carona , a potential candidate for lieutenant governor .**

It also had the backing of Democrats such as Irvine Mayor Larry Agran.

When Chris Matthews noted that **some Democrats were withholding support, Schwarzenegger waved him off , saying , " Let all get together and make this pass and do something for the children . "**

That an appealing message for voters who 've seen enough sharp-edged ideologues.

His accent is another plus .

The GOP anti-immigration image has hurt it among naturalized citizens , a growing force in Orange County .

Who

could better mend this damage than a fellow immigrant ?

Like a Middle European version of Ronald Reagan , Schwarzenegger speaks movingly about what the American dream meant to a young man from a modest home in Austria.

Schwarzenegger **brings Reaganesque excitement and Hollywood glamour**to his public appearances **.**

That a change from the recent run of state Republican leaders.

Can you picture a **George Deukmejian action figure ?**

A Bill Simon video game?

On second thought , the latter is a possibility .

Players would compete to make the most mistakes **.**

Celebrity has drawbacks .

In a run for office , Schwarzenegger would have to answer for the high body count in his movies.

By the way **, Santa Ana MainPlace mall was the filming site for the opening of " Kindergarten**Cop . " (With only one killing , the MainPlace sequence was relatively tame .)

Celebrity also brings out the tabloids **.**

When newspapers reported that Schwarzenegger might run in 2002 , Democratic strategist Garry South provided the press with a magazine article making salacious personal charges.

Some observers say South may have deterred Schwarzenegger from running .

There another possibility .

Maybe Schwarzenegger deliberately raised the possibility of a 2002 race to see **how enemies would go after him in 2006**.

Now that South has revealed the outlines of his attack plan , Schwarzenegger has four years to figure out his responses .

If you doubt that Schwarzenegger is capable of such calculation , see " Pumping Iron ," the 1977 documentary that made him a star (An enhanced version premieres Friday on Cinemax) .

Even as a young bodybuilder , the film shows , he was already waging psychological warfare on opponents **.**

The former Mr. Universe is literally the embodiment of self-discipline .

This fall , he campaigned for his initiative with the same single-mindedness he once applied to his abs and pecs .

He mastered the policy details , stuck to his message and deftly sidestepped questions about other issues .

If Bill Simon had such discipline , he 'd have the governorship – and big biceps too .

Schwarzenegger **also has a knack for making friends in high places**– very high places **.**

Calling Schwarzenegger a " Miracle Man ," the Rev. Robert H. Schullerlet **him promote the initiative in the Crystal Cathedral**, the first such event in the cathedral history .

The actor charm extends to the rank and file .

During the " Hardball " appearance , a **Chapman student posed a probing question about the initiative limitations**.

" First of all ," he responded , " I want to congratulate you for a great question **. "**

Matthews then cracked , " And you 're not a politician , right ?"

Arnold Schwarzenegger is not just the Terminator .

He the Natural

C.7 benetton background

RALEIGH, N.C. –

Dagmar Polzin first saw the man of her dreams in a Benetton fashion ad. Her last vision of him could come this month, through the window of an execution chamber.

Polzin, 32, spotted Bobby Lee Harris in late 1999 on a Hamburg, Germany, bus stop ad. The Italian clothing company used pictures of seven North Carolina death row inmates as part of a controversial ad campaign.

Polzin said she immediately had an instinct about Harris.

By September, she had visited Harris in prison. By October, she moved to North Carolina to be near him.

Polzin and Harris, 34, have asked for permission to marry, a decision that ultimately rests with the prison warden. Time is precious, because Harris is scheduled to die Jan. 19 for the 1991 stabbing death of an Onslow County fisherman – a crime to which he has confessed.

Polzin visits Harris once a week. They have never touched.

"We hold hands through the glass," Polzin said Tuesday in a telephone interview from her west Raleigh home.

"We make the best of things. We are still hopeful. We feel that when you really love somebody, no glass, no people can destroy love."

Harris could not be immediately reached because of prison regulations. In an interview in The Herald-Sun, he said would approach death"like a man."

"I'd sure like to hold my girlfriend's hand and give her a kiss, you know," Harris said.

The state Supreme Court upheld the death sentence after Harris confessed to stabbing fisherman John Redd three times in the back and dumping his body over the side of a boat.

Harris' lawyers said clemency is their primary hope to spare his life. No appeals have been filed.

Gov.-elect Mike Easley will be asked to stop the execution after he is sworn in this weekend, said defense lawyer Mark Edwards.

Edwards will argue that Harris' defense was poor, because one of his lawyers was suffering from cancer.

Jurors were also not informed that Harris has an IQ in the low 70s. The Legislature is expected to consider a bill this month which would bar execution of people with IQs of 70 or below.

If clemency is not granted, Polzin said she will join Harris' family for a contact visit prior to his lethal injection.

She expects to be in the darkened witness room to watch him die.

C.8 benetton new article

A convicted killer who was featured in a fashion company anti-death penalty ads that created an uproar **has been resentenced to life in prison after a lengthy legal battle**.

Bobby Lee Harris **, who once came within hours of being executed , was sentenced Monday by Onslow County Superior Court Judge Paul Jones for the 1991 killing of John Redd**.

Three years ago **, Harris was one of 26 death row inmates**featured in anti-death penalty ads by clothing maker Benetton .

The Italian clothing company used pictures of seven North Carolina death row inmates **and 19 others across the nation the campaign .**

Benetton officials said the ads were meant to raise awareness about the death penalty , but victims ' rights groups said the ads glorified convicted killers and were insensitive to victims ' families and friends.

Harris was convicted of first-degree murder in 1992 , but the jury that was to consider his sentence was dismissed by the trial judge when one of Harris ' lawyers was stricken with bone cancer and was taken off the case.

The judge assigned a new attorney to help represent Harris during sentencing and gave him six weeks to prepare , then called another jury which decided on the death sentence.

Just hours before he was due to be executed in 2001 , Orange County Superior Court Judge Wade Barber vacated the sentence , saying the process had been flawed because different juries judged the trial and sentencing.

Barber also said the trial jury was dismissed without good cause and replaced by a panel that had no authority under state law to hear the penalty phase.

Prosecutors appealed , but the state Supreme Court refused to hear the case.

Defense attorney Chip Medlin welcomed Monday action .

" It was great to finally get a judge to pronounce a life sentence . "

District Attorney Dewey Hudson said the death penalty was appropriate for Harris because of the brutal nature of the crime.

Redd was stabbed during a fishing trip and left for dead in an oyster bed.

After the Benetton ads came out , a German woman , Dagmar Polzin , said she fell in love with Harris when she saw a Benetton ad on a bus stop in Hamburg , Germany **.**

Polzin left her waitressing jobin Germany and moved to North Carolina in 2000 , intending to marry Harris before his scheduled execution **.**

No wedding took place

C.9 bouncer background

Police and a defense attorney offered divergent accounts Tuesday of the nightclub fight that began with a dispute over New York's new smoking ban and ended with the death of a bouncer.

The differing tales came after prosecutors declined to charge two brothers arrested after the early Sunday morning confrontation in the East Village.

A senior police official said a disc jockey saw bouncer Dana Blake pull up short with a startled expression on his face as he pushed stockbroker Jonathan Chan out of the basement of club Guernica around 2:30 a.m.

Blake, 32, died about 11 hours later from a puncture wound to the groin.

Police arrested Chan, 29, and his brother, medical student Ching"Alan" Chan, 31, after the fight on suspicion of assault, criminal weapons possession and resisting arrest.

The police department said early Tuesday that the Manhattan district attorney's office was not prosecuting the brothers and they were being released.

The DA's office was still investigating, spokeswoman Barbara Thompson said.

The Chans' sister, Alice Chan, a Manhattan bookkeeper, also was taken into custody after the fight and was released early Tuesday, said the siblings' attorney, Ivan Fisher.

Fisher said Jonathan Chan left the club on Avenue B before Blake collapsed.

"He did not see what happened," Fisher said. "He certainly did not see Mr. Blake in any form of extremis or injury."

Surgical efforts to patch Blake's wound complicated medical investigators' efforts to determine what killed him, police Capt. James Klein said. But the attending physician reported that Blake's roughly 2-inch-deep groin injury was consistent with a stab wound, Klein said.

Investigators had speculated that Blake may have been injured by a broken bottle, but no glass was found in the wound and the possibility has been largely discounted, police said. No weapon was recovered at the scene, police said.

According to police, Blake approached the men about 2:30 a.m. Sunday to tell them they could not smoke in the bar.

Police spokesman Michael O'Looney said witnesses told police that harsh words were exchanged and the brawl began when Blake tried to eject Jonathan Chan for disorderly behavior. A third man and Alice Chan then intervened, police said.

Ching Chan is a medical student at the Albert Einstein College of Medicine of Yeshiva University in the Bronx.

"These are decent people who do not walk around with weapons in their pockets," their lawyer said. "They live honorable lives, and they're gentle people."

The smoking ban, pushed by Mayor Michael Bloomberg, took effect late last month. It covers all workplaces, including bars and small restaurants. Owners whose establishments violate the ban can be fined or have their licenses suspended.

Tony Blake, the victim's older brother, said he blamed the death on the smoking ban.

"I'm very bitter," he said. "It's a senseless murder because of this stupid cigarette law."

C.10 bouncer new article

The three siblings arrested and then released in the stabbing death of an East Village nightclub bouncer who asked them to comply with the city new anti-smoking law are **the children of a notorious Chinatown gang leaderimprisoned on a federal murder charges**, police said .

Jonathan Chan , 29 and Ching " Allan " Chan , 31 , are the are sons of Wing Yeung Chan , who for a decade was the **leader of the infamous " Ghost Shadows " gang .**

Their sister , Ngan Ling Chan , 33 , was also in the Guernica club on Avenue B early Sunday morning as the brothers feuded with 6-foot-5 bouncer Dana " Shazam " Blake , 32 , who 'd asked one of the brothers to stop smoking .

Manhattan prosecutors **dropped assault charges**against the three on Tuesday **, citing a lack of evidence .**

Their father **, Wing Chan , was president of a Chinese businessman association called the National On Leon Chinese Merchants Assn**.

But his real power was **as street boss**of the Ghost Shadows **, a crime cartel that dealt in extortion gambling and murder , authorities say**.

In the mid-1990s **, Chan was hit with federal racketeering and murder charges - which apparently led him to cooperate with the feds on their effort to fight crime in the Chinese community**.

Originally **charged in several murder cases**, Chan ended up **sentenced to 10 years in federal prison**, law enforcement sources said .

The Chans ' family connections **came to light as police sources reported that a third person probably helped get rid of the knife that killed Blake during**a brawl with the **two blood-drenched brothers**.

Blake **, described by friends as a " gentle giant ,"**was stabbed **fatally in a leg artery**after he told Jonathan Chan to put out his cigarette **.**

The Chans , whose clothes and shoes were soaked with blood when they were arrested, were arrested but later freed after the Manhattan district attorney office declined to prosecute .

The weapon has not been found **.**

Police sources said a pile of Ngan Ling Chan bloody clothes was found in her Chinatown apartment.

She explained the find by telling cops it was her own menstrual blood.

Ivan Fisher , a lawyer for the Chans , said his clients are innocent **.**

" Police and prosecutors are under intense pressure to charge them for something they did not do .

They are hoping that reason and conscience and professionalism will prevail. "

Meanwhile **, cops said they were interviewing cab drivers and re-interviewing witnesses at the bar**, while **investigators combed through sewers and garbage cans in search of the weapon**.

Blake family members said they were outraged that the Chans were free **.**

" These guys are walking around smelling fresh air , and my brother is dead," said Blake brother Anthony **, an associate minister**at **the Humble Way Church in Astoria ,**Queens **.**

" We will stay with the case until someone is brought to justice. "

" It an insult to say there not enough evidence here ," said the Rev. Mitchell Taylor , the Blake family minister .

At a rally yesterday in the Astoria housing complex where he lived , Blake friends and family demanded justice

---

C.11 brando background

Marlon Brando has settled a $100 million breach of contract suit brought by a former maid who is the mother of three of his children, attorneys for both sides said Wednesday.

"It has been amicably resolved," said Donald Woldman, as he left the courthouse with his client, Maria Cristina Ruiz, 43.

Ruiz declined to comment and hid her face from photographers with a scarf as she left the courtroom. The 79-year-old actor did not attend the hearing, which was closed to the public.

No details of the settlement were revealed.

"It's a private matter and it should be left private," said Leon F. Bennett, Brando's attorney.

Ruiz once worked as Brando's maid and lived at his home after they became romantically involved in 1988, according to the suit, filed in Superior Court in April 2002.

She said the relationship ended in December when the actor stopped paying her living expenses, although he continued to support her children, who are 8, 10 and 13.

Brando, who has been married three times, has nine children.

In her lawsuit, Ruiz said she"devoted all aspects of her life to... Brando's needs, the interests of their children, his personal interests and well being, to the exclusion of her own."

In return, the lawsuit said, the Oscar-winning actor"promised that he would always provide for and financially support plaintiff and any children of plaintiff and defendant Brando."

"It was just like a marriage. There was an engagement ring and a wedding ring," Woldman said when the suit was filed. He noted, however, that there was never a legal marriage ceremony.

Brando had disputed Ruiz's claims, dismissing their relationship as"nothing more than sexual.

The actor, who won Academy Awards for 1954's"On the Waterfront" and 1972's"The Godfather," had also said he was broke despite an income of nearly $8,000 a month. He blamed mounting legal bills, adding he had been forced to mortgage his Los Angeles home.

C.12 brando new article

A $100 million breach of contract suit against Marlon Brando by a former maid who bore three of his children **will be dismissed soon**, attorneys for both sides **informed a judge today**.

Attorneys **told Los Angeles Superior Court Judge Roy L. Paul**that they planned to ask for dismissal of Maria Cristina Ruiz civil suit .

It **was not immediately clear , however , when the dismissal would be final**.

After the lawyers made their announcement in court , the question of **the custody of the couple three children still remained to be resolved**.

At that point **, the judge asked the media to leave the proceeding because a " confidential filing " was involved.**

That case is sealed , meaning no court documents or future court dates are made public.

Ruiz sued last April, claiming she had a 13-year relationshipwith the big screen legend and gave birth to three of his children , now 13 , 10 and 8 , respectively .

Ruiz , 43 , claims the 79-year-old star of " The Godfather **" promised to take care of her and their children until the relationship ended**, then **to "**

divide equally any income and property acquired " during the pairing, which began in February 1988 and ended in December 2001 , according to the suit .

Ruiz attorney Donald Woldman has called the relationship " just like a marriage , except without the ring . "

Brando has been **married four times and an online biography reports he has seven children .**

Ruiz , who **wore a black , hooded sweater in court**, claims Brando covered her living expenses until the end of 2001 , then dropped her from the picture .

She was seeking $100 million in damages and an order requiring the two-time Academy Award winner **to pay " a reasonable sum " each month**.

Brando was not present for the hearing **, but reportedly was in the courthouse earlier**.

The method actor , who **shot to stardom with the portrayal of washed-up boxer Terry Malloy**in " On the Waterfront ," claimed that despite a monthly income of almost $8,000 he is broke because of mounting attorney fees .

He said he has had to mortgage his Los Angeles home .

When the suit was filed last year **, Brando called it " gutter**served ," and that his relationship with Ruiz was " nothing more than sexual **. "**

The reclusive Omaha , Neb. , native has made six films since " The Godfather , Part II " in 1977

C.13 bubonic background

A professor at the Texas Tech University Health Science Center was arrested Wednesday night and accused of giving false information to the FBI about 30 vials of plague he reported missing at the university.

U.S. Attorney Dick Baker said Dr. Thomas C. Butler was arrested on a complaint of false statement to a federal agent. Baker said Butler said the vials were missing as of Jan. 11 when"truth in fact, as he well knew, he had destroyed them prior to that."

University spokeswoman Cindy Rugeley said Butler, the project's principal investigator, reported the vials as missing.

Butler has been at Texas Tech since 1987. He is chief of the infectious diseases division of the department of internal medicine. The university said he has been involved in plague research for more than 25 years and is internationally recognized in the field.

Butler was booked into the Lubbock County Jail about 8 p.m. and will be arraigned by a magistrate on Thursday.

The report of missing vials triggered a terrorism-alert plan and showed how jittery Americans are over the threat of a biological attack.

The samples, about 30 of the 180 the school was using for research on the treatment of plague, were reported missing to campus police Tuesday night.

"We have accounted for all those missing vials and we have determined that there is no danger to public safety whatsoever," Lubbock FBI Lupe Gonzalez said earlier Wednesday.

Baker said FBI agents interviewed Butler on Tuesday. He said the complaint pointed out that the false statement resulted in a huge investigation involving about 60 state, local and federal agents.

The public did not learn of the report of missing vials until early Wednesday but hospitals and medical personnel were notified Tuesday, part of the city's post-Sept. 11 emergency plan.

"We didn't want to spread panic," said Tech Chancellor David Smith. "As it turns out, they were never missing." He would not elaborate.

Smith declined to comment, through Rugeley, about the arrest.

Mayor Marc McDougal said the public was not notified because of information the university received late Tuesday that indicated the missing vials were not a threat to the public.

"I think when you look how quickly it came down and how it got resolved, I think it would be hard to second guess" how we handled it, he said. "One thing we didn't want to do was cause people to panic."

The vials were kept in a secure area that does not have a surveillance camera and that there is limited access to the area, officials said.

"I don't know the precise number (of keys), but it's limited," Smith said. "Policy (for federal grants) was not violated. This is one where we're looking at the human element."

Plague - along with anthrax, smallpox and a few other deadly agents - is on a watch list distributed by the government, which wants to make sure doctors and hospitals recognize a biological attack quickly.

Health officials say 10 to 20 people in the United States contract plague each year, usually through infected fleas or rodents. The plague can be treated with antibiotics, but about one in seven U.S. cases is fatal.

Texas Tech said that officials thought it was"prudent" to get law enforcement involved because of current concerns about bioterrorism.

The FBI sent agents to Lubbock. The Centers for Disease Control and Prevention also took part in the investigation. Homeland Security chief Tom Ridge contacted McDougal. About 60 investigators from the FBI and other agencies converged on the medical school Tuesday night.

The form of the disease called bubonic plague is not contagious. But left untreated, it can

transform into the more dangerous pneumonic plague that can be spread person to person. The most infamous plague outbreak began in 1347 and killed 38 million people in Europe and Asia within five years.

EDITORS - Associated Press Writer Curt Anderson in Washington contributed to this report.

CDC:URL

C.14 bubonic new article

A federal grand jury issued a 15-count indictment Thursday charging a Texas Tech University professor with falsely reporting that vials of bubonic plague bacteria were missing from a laboratory **.**

Thomas C. Butler also was charged with smuggling plague bacteria into the United States, illegally transporting the bacteria to a lab in Fort Collins, Colo. , and overseas **, lying to federal agents and filing a false income tax return**.

The scientist reported in January that 30 vials of the bacteria responsible for the deadly disease were missing from a university lab .

The report , coming amid public worry over biological attack , triggered a terrorism-alert plan and prompted the FBI to rush dozens of agents to the West Texas city .

Butler , 61 , a noted researcher **who had previously worked at Johns Hopkins and Case Western Reserve universities, is on paid leave from Texas Tech medical school**, where he is chief of the infectious diseases division .

Butler attorney , Floyd Holder **, said Butler will plead innocent to the charges.**

" Since the reason for all these activities was to aid this nation in its effort to defend against bioterrorism , it seems cruel thanks ," Holder said .

" I 'm disappointed that the government felt they had to do this to one of the nation top research scientists. "

U.S. Attorney Jane Boyle said in a statement that the " incident could have sparked widespread panic **. "**

" This case is an excellent example of how , in the present climate , authorities at all levels of government are approaching their commitment to protect the public with cool heads and joined hearts ," she said.

According to court documents **, Butler was the only person with authorized access to the lab bacteria.**

Butler gave a handwritten statement to the FBIsaying **he had accidentally destroyed the bacteria and made a " misjudgment " by telling school authorities that the vials were missing**, the documents indicated .

Butler was charged with lying to FBI agents by reporting vials as missing when he knew they had already been destroyed .

The indictment also alleges that Butler **brought plague bacteria samples on a plane from Tanzania to Lubbock in April 2002**and **did not fill out paperwork disclosing the samples.**

He was charged with improperly driving samples to a Centers for Disease Control facility in Fort Collins , Colo. , shipping 30 vials to Tanzania via FedEx, and **sending others aboard an American Airlines flight to a U.S. Army research center in Fort Detrick ,**Md.

Some of the samples **were marked simply as " laboratory**materials ," according to the indictment .

Butler lawyer has **said the scientist secured the samples in a plastic container in his luggage.**

The lawyer said Butler used the bacteria for preparatory work in seeking a $700,000 federal grant to study treatment of the plague .

The indictment charged that Butler falsely told FBI agents he did n't know the legal requirements for transferring plague bacteria.

The grand jury **also charged Butler with filing a false income-tax return for 2001 by listing $114,000 in payments from two pharmaceutical companies as a business expense , which reduced his taxes by nearly $40,000.**

If convicted on all counts , Butler could face up to 74 years in prison and $3.6 million in fines, authorities said .

Bubonic plague is an infectious disease usually spread to humans **by handling an infected animal or being bitten by a rodent flea**.

Outbreaks that led to massive death tolls were common in the Middle Ages .

Along with anthrax , smallpox and a handful of other deadly agents , bubonic plague bacteria is on a federal watch list designed to help doctors and hospitals quickly recognize a bioterror attack

---

C.15 celebs background

A Los Angeles police officer is under investigation on charges he used an LAPD computer system to gather and sell private information about Hollywood stars to a tabloid.

Pending the outcome of the probe, officer Kelly Chrisman has been placed on"home duty" by L. A. police Chief William J. Bratton, former Boston and New York police commissioner.

Jennifer Aniston, Sean Penn, Sharon Stone, Meg Ryan, Kobe Bryant, O. J. Simpson, Larry King, Cindy Crawford, Drew Barrymore and Farrah Fawcett reportedly are some of the names included on Chrisman's celebrity tracking list.

According to the L. A. Times, the probe was launched after Chrisman's ex-girlfriend, Cyndy Truhan, filed suit against the city, claiming he had"used his position as an officer to find her new telephone numbers and addresses," following their break-up.

But before the City Council settled the suit for $387,500, she alleged he had boasted he had gathered confidential law enforcement records about celebrities and sold the information to the National Enquirer. The tabloid denied purchasing information from any police officer.

The Times also reported"numerous" phone calls were traced from Chrisman to the tabloid.

Chrisman, 34, admits collecting the data, but claims he did so on orders from superior officers who were assembling a map of VIP residences for more efficient law enforcement, according to the Times.

"There's really nothing in those records to sell to tabloids," said Chrisman's attorney, Christopher Darden. "He didn't do that."

Regarding Chrisman's claim he was acting on orders, police spokesman officer Jack Richter told the Herald yesterday, "Whether or not that's true or not is. . . under investigation right now."

Meanwhile, City Councilor Dennis Zine told the Herald yesterday, "We're talking about other individuals whose information was run that could be potential additional liability. . . . The police department needs stronger auditing to make sure this kind of (alleged) thing doesn't happen."

C.16 celebs new article

A Los Angeles police officer used department computers to access confidential law enforcement records of celebrities and sold the information to tabloids **, according to a lawsuit recently settled by the city .**

Officer Kelly Chrisman **, a 13-year veteran**, acknowledged looking up the information , but said he did so at the direction of his superiors , according to internal Los Angeles Police Department records .

Attorney Christopher Darden said his client never sold the information to anyone .

The lawsuit prompted the department to launch its own investigation , which the Los Angeles Times reported Tuesday **turned up " hundreds of hits " on the names of famous people**, including Jennifer Aniston **, Pamela Anderson , Sharon Stone**, Sean Penn , Meg Ryan , Kobe Bryant and **Nicole Brown Simpson .**

Personal information available through the department computer system goes far beyond background on those with criminal records .

Officers can access government agency files on individuals ' driving records , birth dates , ownership of vehicles , physical descriptions , Social Security numbers , restraining orders and , in some cases , unlisted phone numbers .

The lawsuit was filed by Chrisman former girlfriend **, Cyndy Truhan , who is also the ex-wife of former Los Angeles Dodgers star Steve Garvey**.

In March, the city paid $387,500 to settle the suit .

Police investigators **said they could not confirm**the 34-year-old officer collected the data for personal financial gain .

Chrisman was placed on home duty , similar to paid leave , while the allegations are investigated

---

C.17 charities background

The new cases bring the number of people charged with 9/11 fraud to 245 - and the amount stolen to $3.6 million, according to prosecutors.

More than half of the latest batch of defendants live in a homeless shelter on the fringe of Chinatown.

Officials said the suspects got thousands of dollars by claiming that the attack left them homeless - but they already were homeless when the twin towers fell.

Manhattan District Attorney Robert Morgenthau said that one homeless married couple stole $6,607 from the Red Cross by providing each other with forged landlord notes.

Another married couple living in the shelter took the largest amount, $25,895. They claimed that their apartment and possessions were totally lost.

Steve Pasichow, deputy commissioner of the city Investigation Department, said that after some at the shelter bragged about what they had done, a worker there alerted a supervisor, who called city investigators.

In another case, a 911 police operator got $9,366 from the Federal Emergency Management Agency after being out of work almost three months for post-traumatic stress disorder. Officials said she actually was home caring for her son, who had broken his leg.

Morgenthau said the theft of relief funds has mostly not involved large rings of crooks but individuals who hear of fraud opportunities"by word of mouth."

"The cases are important not because there's a huge amount of money involved but because. . . people [must] understand you cannot take advantage of a national tragedy," Morgenthau said.

C.18 charities new article

A continuing crackdown on people who have attempted to defraud federal relief programs and private charities related to the 9/11 attack has **resulted in the recent arrests of 33 people, who together falsely claimed nearly $135,000 in government and charitable funds**, prosecutors said yesterday .

The arrests bring to 245 the number of people charged with crimes related to World Trade Center fraud , said Robert M. Morgenthau , the Manhattan district attorney .

Those cases resulted in about $3.6 million in fraudulent claims .

" These cases are important not because such a huge amount of money is involved ," Mr. Morgenthau said , " but because we want to make sure that people understand you ca n't take advantage of a national tragedy . "

Among the arrests announced yesterday **were 16 former**residents of a downtown homeless shelter who claimed that they were forced out of their homes by the Sept. 11 attack , officials said .

In fact , Mr. Morgenthau said **, each of the 16 were already living at the LIFE Shelter at 78 Catherine Street , before Sept. 11**.

Steven Pasichow , the inspector general for the New York City Housing Authority , said an employee of the shelter tipped off the authorities .

The shelter residents collected **benefits ranging from less than $500 to nearly $26,000**.

The district attorney **also announced the arrest of 11 people , including a 911 operator for the Police Department and two school crossing guards**, who were charged **with falsely applying for relief benefits by saying that they had lost jobs because of the terror attack**

C.19 dahlia background

A retired homicide detective believes that his father committed the notorious"Black Dahlia" murder that has been unsolved for half a century.

In a new book, Steve Hodel states that late physician George Hodel killed 22-year-old aspiring actress Elizabeth Short in a fit of jealousy.

Her body, severed at the waist, nearly drained of blood and posed with arms and legs spread-eagle, was found in a vacant lot in 1947. Authorities said her face and body had been slashed, apparently while she was alive.

Short was nicknamed the"Black Dahlia" by acquaintances because of the black clothing and the flower she wore in her hair.

In"Black Dahlia Avenger: The True Story," the former Los Angeles police detective also claims that his father killed another woman, Jeanne French, less than month later. That killing was dubbed the"Red Lipstick Murder" because the cosmetic was used to scrawl"B.D." on the body - a possible reference to the Black Dahlia case.

"The man responsible for these terrible murders, that of the Black Dahlia and the Red Lipstick Murder, and most probably many others, is my own father, Dr. George Hill Hodel," Steve Hodel told a press conference Friday.

"I'm asking the press and the detectives to now take up that responsibility and reopen the murder books and continue the investigations so that all of us may know the truth," he said.

He believes his father was a serial killer who may have committed 20 unsolved murders in the 1940s and 1950s.

There have been many other claims by authors to have discovered the Black Dahlia's killer. The Los Angeles County district attorney's office was aware of Steve Hodel's claims but noted that even if they were true, his father's death leaves no one to prosecute.

Steve Hodel said he began his investigation of his own father by chance. He said he looked through his belongings after the man died in 1999 and found two photographs of a woman that he now believes is Short.

"I've been to hell for the last three years," the son said.

However, Detective Brian Carr, who oversees the Police Department files of the Black Dahlia case, said he was unable to determine whether they were photographs of the victim.

Hodel also alleged similarities in notes written by his father and one the killer supposedly sent to a newspaper after the murder.

C.20 dahlia new article

It a story that fascinated crime buffs for decades .

There have been almost as many potential murderers put forward as fictionalized versions of the long-unsolved case .

Now , a onetime Van Nuys cop has penned a book that names the latest suspect in 1947 grisly " Black Dahlia " mutilation slaying - his own father .

Author Steve Hodel claims his physician father , the late George Hodel , was a serial killer **who probably killed 20 women**, including Elizabeth Short , the 22-year-old Hollywood hopeful whose mutilated and horribly posed body was found in a vacant lot **near Leimert Park on Jan. 15,**1947 .

Dubbed " the Black Dahlia " because of her black clothing and the flower she often wore in her hair , Short slaying **prompted the largest manhunt in Los Angeles police history and remained on the front pages for weeks .**

" I 'm asking the press and the detectives to now take up that responsibility and reopen the murder books and continue the investigations so that all of us may know the truth ," Hodel said

Friday .

Now a private investigator , Hodel pins about 20 other killings of women in the 1940s and 1950s on his father **, who died in 1991at age 91 and was never arrested in any of the crimes.**

A musical prodigy and onetime crime reporter, the late **Hodel was a doctor specializing in venereal disease control for the county health departmentand opened VD clinics that treated well-connected Los Angeles residents.**

The author claims his father , the doctor , kept dossiers on clients.

In the just-published " Black Dahlia Avenger : The True Story ," Hodel says his father **had become a primary suspect in the Dahlia slaying and was about to be arrested when the case was suddenly dropped because authorities feared Hodel had friends in high places and would leak the medical files of well-known patients , including police officials.**

" At that point , the whole thing was shut down and my father left the country for 40 years and everything sunk below the surface ," said Hodel, who **spent nearly two dozen years with the Los Angeles Police Department**, including on Van Nuys patrol in the 1960s before **working his way up to homicide detective in Hollywood**.

Hodel story of slaying , mystery and corruption **suggests a true-crime " Mommie**Dearest " as well as Geraldo Rivera televised Al Capone vault fiasco **.**

He admits there is not one person alive who can corroborate his allegations, including claims that a corrupt 1949 grand jury **was ordered to stop hearing evidence because of fears the LAPD reputation would become tarnished if the VD files became public .**

In the book , Hodel says his father killed Short and **sliced her body in half with a surgeon skill during**a fit of jealousy .

Weeks later , the nude , beaten corpse of 45-year-old Jeanne French was found .

Dubbed " the Red Lipstick Murder " because the initials B.D. (suggesting a connection to the Black Dahlia) were scrawled on the body in red lipstick , Hodel claims that killing , too , was his father work **.**

Stephen Kay , a Los Angeles County deputy district attorney who helped prosecute the Manson clan inthe 1970s **, said in a statement he would n't have hesitated to file murder counts against the elder Hodel in connection with the two slayings.**

He said he had " no doubt " George Hodel " not only murdered Elizabeth Short (the Black Dahlia) but also murdered Jeanne French. "

Ultimately , Hodel " Black Dahlia Avenger " offers merely the latest theory in a crime that continues to capture the imagination .

Books - **including crime writer James Ellroy fictionalized " The Black Dahlia " as well as the 1981 film " True**Confessions " - borrowed details of the case .

And in a 1995 paperback **, writer Janice Knowlton claimed her father was the Black Dahlia slayer.**

The 61-year-old Hodel , who lives in Lake Arrowhead where he runs a private detective agency , told reporters he began investigating his father - who abandoned the family when he was 9 years old- after finding photographs among his late father possessions of a woman he believes is Short .

But a detective who oversees files of the case , said it was impossible to determine whether the photos were actually of Short .

Hodel also saw similarities in notes the killer supposedly sent to newspapers and his father handwriting

---

C.21 drchaos background

A man who called himself"Dr. Chaos" and vandalized power lines and transmission towers in Wisconsin was sentenced Thursday to 13 years in prison for hiding deadly cyanide in a Chicago subway tunnel.

Joseph Konopka, 26, a former computer systems administrator from De Pere, Wis., was at a loss when U.S. District Judge Wayne Andersen asked why he had gone on his vandalism spree.

Konopka also pleaded guilty December 20 to six federal law violations related to conspiracy to destroy energy facilities, arson of buildings, trafficking in counterfeit goods, intercepting electronic communications and damaging a protected computer for a Wisconsin crime spree.

Konopka is expected to receive a 20-year sentence for those crimes in federal court. Sentencing was scheduled April 11.

Andersen said that sentence was likely to be consecutive to his 13-year term for the two bottles of cyanide that he took from an abandoned chemical warehouse and hid in a Chicago Transit Authority subway tunnel.

Konopka was arrested in March 2002 by police who had staked out a tunnel under the University of Illinois-Chicago, where a 15-year-old had been caught one night earlier; the teen told police a man was storing cyanide. Konopka later led police to the two large bottles of cyanide hidden in the subway tunnel.

Defense attorney Matthew Madden told the judge Konopka found the cyanide during his"urban exploration," which involved nightly walks through dark tunnels and abandoned buildings.

Madden said his client's penchant for destruction and bizarre behavior"stems from an abnormal maturation process." He said normal adults"realize you can't participate in the destruction of property for your own entertainment - that's just not acceptable."

Konopka said that when he found the cyanide in a South Side warehouse"I made the bad decision of grabbing several bottles and packaging them up and taking them."

He also said he had considered using the cyanide to commit suicide but never intended to hurt anyone.

"I certainly apologize to the city of Chicago, the citizens and all of the people who had to retrieve the chemicals," he said.

The judge said he could tell from Konopka's apology that he was smart and"you know right from wrong." But he included in the sentence a warning to the Bureau of Prisons to beware of anything dangerous that Konopka might do involving computers and chemicals while in federal custody.

C.22 drchaos new article

The 26-year-old man who called himself " Dr. Chaos " and sent emergency and CTA workers into a panic last year over his stash of cyanide was sentenced to 13 years in prison Thursday **under a never-before-used law**.

Stephen Konopka of DePere , Wis. **, was the first person to be charged with possessing a chemical weapon in the decade the statute has been on the books**, according to federal prosecutor David Weisman .

" The conduct jeopardized the safety of the public , and in this case it appropriate to fully prosecute ," Weisman said , noting the potential for harm .

An alleged computer hacker , Konopka was arrested last March while fleeing an underground tunnel in a University of Illinois at Chicago building **.**

When police discovered cyanide on Konopka, he led them to his stash in a **CTA substation near the Washington-Dearborn stop on the Blue Line , according to Konopka plea agreement last November .**

Konopka and a 15-year-old boy took the powdered cyanide and other chemicals from an abandoned warehouse near **48th and Halsted , according to court records .**

Konopka also faces a 20-year term for damaging communication and energy power stations near **Green Bay , Wis. , leading to disruptions for more than 30,000 customers**.

He will be sentenced under a plea agreement in Milwaukee on April 11 .

A judge there will determine whether the two sentences will be served consecutively **.**

" A lot of it started when I was young ," a somber but cooperative Konopka told U.S. District Judge Wayne Andersen when asked to explain his actions .

It was " a pattern , and I never got out of it ," he said .

After Konopka arrest last year , his relatives told reporters that he had suffered from depression and a history of suicide attempts stemming from a rough childhood , including his mother kicking him out .

Before sentencing , Andersen cited the potential danger of Konopka actions **.**

Konopka attorney , Matthew Madden , said " it pretty obvious that Mr. Konopka is not an angel , and we 're not trying to say that he is ."

Konopka apologized to the city , the CTA and " people who endangered themselves to retrieve these chemicals **. "**

In response , Andersen said he did not feel Konopka acts were " mean-spirited . "

" There mental health issues here , and I do n't think that something to be avoided or to be ashamed of ," Andersen said .

Konopka could have received 11 to 14 years under federal sentencing guidelines **, but Andersen fell in between Weisman request for the toughest sentence and Madden plea for the lightest punishment .**

He said Konopka mental issues were a factor in him not giving the harshest sentence possible .

Andersen stressed that Konopka should receive drug treatment and mental health counseling while in prison.

He also will warn the Bureau of Prisons to be cautious about Konopka access to chemicals and to monitor his computer use

---

C.23 elijah background

More indictments charging teen with threatening president

Additional charges have been included in a new indictment handed up against a young man accused of threatening to kill the president.

Elijah Wallace, 18, of Brentwood, was charged with three counts of threatening the life of the president and two counts of threatening others, the U.S. attorney's office said Thursday.

Each charge carries a term of five years in prison.

Wallace originally was charged with one count of threatening the president in an indictment returned in March.

Wallace had pleaded innocent to threatening to kill the president.

The government said Wallace made the threats in letters he mailed from jail to WMUR-TV in Manchester and a friend, and in statements he made to others. He was being held on a burglary charge at the time.

In a Feb. 16 letter to the television station, he expressed a desire to shoot the president with a rifle at point blank range, the U.S. attorney's office said. Additional statements included"everyone in this pathetic government must be executed," the office said.

When arrested in January on that charge, authorities said Wallace told them he had sent letters containing anthrax to Senate Majority Leader Tom Daschle in Washington, the state Motor Vehicles office in Concord and two businesses.

Investigators said it was unlikely Wallace sent the letters. A white powder found in his possession when he was arrested tested negative for anthrax.

Wallace's father, Eric Wallace, said his son is a compulsive liar who probably made up the letters to get attention and to make himself feel important.

He said his son has a high-functioning form of autism and has been in and out of jail in recent years. He said the teen told him he would like to get into counseling to get his life back on track.

C.24 elijah new article

Teen pleads guilty to threatening President Bush

A teenager has pleaded guiltyto sending a letter to a television station in which he threatened to shoot President Bush **.**

Elijah Wallace , 19 , of Brentwood , admitted sending the letter to WMUR-TV in February 2002, in which he expressed a desire to shoot Bush with a rifle at point-blank range .

The letter contained additional statements such as , " everyone in this pathetic government must be executed . "

Wallace appeared in U.S. District Court on Tuesday .

A sentencing hearing was scheduled for June 17.

The charge carries a term of up to five years in prison .

Wallace also told police he had sent letters containing anthrax to Senate Majority Leader Tom Daschle in Washington , the state Motor Vehicles office in Concord and two businesses **.**

Although Daschle office did receive a letter containing real anthrax in 2001 , Wallace was quickly ruled out as a suspect .

The other offices did not see such letters .

Wallace currently is serving a two-year sentence in state prison .

Last year , he pleaded guilty in a New Hampshire court to burglary and other charges after he broke into a home and told police he was mailing anthrax-laced letters

to politicians

C.25 freeway background

A single engine Cessna clipped a car and shattered its rear window as the plane landed on the Riverside Freeway Saturday afternoon, authorities said. No one was hurt.

Pilot Mike Manning, of West Covina, was going to land at nearby Fullerton Municipal Airport but had engine trouble and landed about 1:25 p.m., said Highway Patrol Officer Colleen Richardson.

Manning's plane clipped a blue Honda Accord driven in light traffic by a 30-year-old Moreno Valley man who was taking his girlfriend and their two children to Disneyland.

Kevin Glovinsky, 40, said he saw a shadow over his car and heard girlfriend, Tiffany Jennings, 32, screaming that an airplane was landing on the freeway.

The plane's wheel smashed into the car's rear window, narrowly missing Glovinsky's 4-year-old son, Jonathan, and Jennings' 8-year-old daughter, Brittany, who was asleep on the back seat.

"I'm still kinda in shock," Glovinsky told The Orange County Register. He said the family canceled their trip to Disneyland and went home to wash broken glass out of their children's hair.

The Cessna 172 landed safely seconds later near the Tustin Avenue exit and pulled to the shoulder, shutting down two westbound lanes for about three hours, Richardson said.

Manning, 45, and his passenger, Michael Aguilar, 48, of Pomona, said they were glad to be alive.

"I couldn't believe it was happening," Manning told the paper. "I'm still in a period of denial."

C.26 freeway new article

A pilot test flying a small aircraft made an emergency landing on an Orange County highway Sunday night after experiencing engine trouble .

No injuries were reported .

It was the second emergency plane landing in eight days .

Pilot Robert Atkins , 41 , was flying to Palomar Municipal Airport about 7 p.m. , when his plane lost oil and blew a piston , said Aaron Reich , a California Highway Patrol dispatcher .

Atkins landed his Cherokee 180 on Ortega Highway at Greenfield Drive .

Traffic was light at the time , and the plane was towed away by a flatbed truck , Reich said .

Atkins was testing an experimental engine when the trouble occurred , he said .

The Federal Aviation Administration and the National Transportation Safety Board were not investigating Sunday emergency landing because no crash had occurred , said Reich.

On April 5, the pilot of a single engine Cessna clipped a car and shattered its rear window as the plane landed on the Riverside Freeway in Anaheim .

The pilot also had experienced engine trouble and tried to land at nearby Fullerton Municipal Airport .

No one was hurt in that landing

C.27 harrington background

In an unprecedented move, Iowa Gov. Tom Vilsack is considering a reprieve for a man whose 1978 murder conviction was overturned two months ago.

The Iowa Board of Parole will meet Thursday to recommend to Vilsack whether 44-year-old Terry Harrington should be released from prison.

A reprieve would essentially suspend Harrington's sentence but place parole conditions on his release until legal issues are resolved. If a reprieve is granted, it will be the first that anyone can recall in Iowa.

"Harrington's case is unprecedented because (the governor) is asking for a reprieve," said Clarence Key, executive director for the Iowa Board of Parole. "To my knowledge, there has never been a case like this go to the board. ... No one around here can remember a reprieve."

Harrington was one of two Omaha teenagers convicted in 1978 for the shotgun murder of John Schweer, who was a retired Council Bluffs police captain working security at an auto dealership in July 1977 when he was gunned down.

Potential appeals appeared at an end for Harrington until a friend got records on Harrington's case from the Council Bluffs Police Department in 1996. The records had information about other suspects in Schweer's murder that was not turned over to Harrington's defense attorneys.

The Iowa Supreme Court ruled in February that the withheld evidence hurt Harrington's defense.

Despite a ruling by the high court, Harrington remains incarcerated at the Clarinda Correctional Facility in southwest Iowa - even though he was granted a new trial.

Vilsack's office said Harrington is in"bureaucratic limbo for an indefinite period of time" because the Iowa Attorney General's Office has asked the high court to clarify its decision. The attorney general's request won't overturn the ruling, but it has delayed the Supreme Court from issuing its order vacating the conviction.

In a letter to the Iowa Board of Parole, Vilsack's office asked the panel to recommend whether he should grant Harrington executive clemency in the form of a reprieve.

"What it's doing is giving Mr. Harrington a chance to walk free, and he's ecstatic," said Anne Danaher, the woman who uncovered the police records and has championed Harrington's cause.

Vilsack's staff has been inundated with letters from Harrington's supporters for more than a year. By last summer, Vilsack's clemency file on Harrington had nearly 200 letters from family members, friends and advocates pleading with Vilsack to pardon Harrington or reduce his sentence.

Even if Vilsack grants a reprieve, that won't stop the Pottawattamie County Attorney's Office from filing new charges against Harrington in the case.

County Attorney Matt Wilber and the Council Bluffs Police Department have reopened an investigation into the murder.

Harrington's mother, Josephine James, said Vilsack's clemency request is a first step, but she will feel better once she knows her son has been exonerated.

"I'm not satisfied with it, but if we can get him out, the Lord will take care of the rest," said James, who was busy cleaning Monday in hopes that her son would be home this week.

C.28 harrington new article

A man who spent nearly 26 years in prison on a murder conviction that was overturned earlier this year **walked out of the state prison here Thursday after Gov. Tom Vilsack signed a reprieve .**

Terry Harrington mother , sisters and other family members met him at the Clarinda Correctional Facility and had a Humvee stretch limousine waiting at a nearby motel to take him to the family home in Omaha , Neb. , and a welcome dinner at his mother church .

He walked out of the prison building at 2 p.m. to screams of joy from his mother , Josephine James , and his daughter , Nicole Brown , 25 , who took his hands and walked with him out of the prison gate .

" Oh God , it good ," Harrington said , stepping on the other side .

" It so good .

Thank you , Jesus . "

He thanked his lawyers , the parole board and the governor for arranging the reprieve .

" As far as we know , this is the first time this has ever been done in Iowa ," said Fred Scaletta , spokesman for the Corrections Department .

" Sometimes it gets done in states with capital punishment where a governor offers a reprieve if someone is going to be executed , but it can also apply in this type of case. "

Harrington , 44 , was just a teenager when he was convicted in the 1977 shooting death of John Schweer , a retired Council Bluffs police officer working as a security guard at a local car dealership .

The Iowa Supreme Court overturned Harrington conviction in February, based on new evidence that prosecutors withheld police reports pointing to another suspect and **that the state key witness had recanted his testimony .**

In such cases , inmates usually are freed while prosecutors determine whether to seek a new trial .

After Vilsack learned this month that Harrington was still in prison because state attorneys were arguing over legal language in the ruling that was n't substantive to the case , he asked the Iowa Board of Parole to review the case and suggest an appropriate remedy .

" Mr. Harrington has seemingly been placed in bureaucratic limbo for an indefinite period of time ," the governor wrote **.**

The Parole Board met early Thursday and recommended a reprieve , which the governor signed at 11:50 a.m.

Harrington sobbed as he was reunited with his family outside the prison .

With tears streaming down his face , he told reporters " I did not commit the murder of John Schweer ... The crucial evidence that Council Bluffs authorities withheld and deliberately conspired to camouflage during all my previous court proceedings have only gone to promote that lie . "

Pottawattamie County Attorney Matthew D. Wilber said he was disappointed that

the governor chose to interject himself into the murder case .

" Until the order comes down from the Iowa Supreme Court , Terry Harrington remains a convicted murder ," Wilber said in a statement .

Once the order is issued , he said , Harrington will be placed in county custody and will have to apply for release on bond pending trial .

" At this time , barring any unforeseen changes in the case , it is our intention to put Mr. Harrington back on trial ... "

Wilber said **.**

Harrington attorney , Thomas Frerichs , said if Wilber decides to seek another trial , " we are fully prepared to prove Terry innocent . "

Bridget Chambers , assistant attorney general , said the matter of law that put Harrington case in limbo could be very important for future appeals cases .

The Supreme Court ruling said that an appeal filed after three years must be based on new evidence that could not have been discovered before , but that it need not be evidence that exonerates the defendant .

If the language stands , it could open the floodgates , clogging courts with time-consuming appeals that might not win the defendant release , she said .

" It will also provide false hope for people who as a matter of law cannot win ," she said .

" That why I considered long and hard to file the petition for review ," Chambers said .

" I knew it would delay Harrington appeal .

But there really is a really good reason here ."

Associated Press Writers Dave Pitt and Melanie Welte contributed to this report

---

C.29 heart background

CHICAGO, April 1 (UPI) – A six-month ban on smoking in public places in Helena, Mont., resulted in a 60 percent reduction in heart attack admissions to the local hospital, researchers said Tuesday. "This striking finding suggests that protecting people from the toxins in secondhand smoke not only makes life more pleasant – it immediately starts saving lives," said Stanton Glantz, professor of medicine at the University of California, San Francisco cardiovascular research institute and a statistics authority. "This work substantially raises the stakes in debates over enacting and protecting smoke-free ordinances," he added.

From June 2002 until December 2002, an ordinance in the city of Helena, Montana's state capital, banned smoking in bars, restaurants, casinos and workplaces in the city. Within a couple of months, Sargent said, heart attacks became rare.

C.30 heart new article

CHICAGO –

Heart attacks in Helena , Mont. , fell by more than half last summer , after voters passed a broad indoor smoking ban , suggesting that cleaning up the air in bars and restaurants quickly improves health for everyone , **a study has found.**

Doctors said their study is the first to examine what happens to public health when people stop smoking and breathing secondhand smoke in public places .

The doctors , themselves backers of the ban , acknowledged that such effects need to be demonstrated in a larger locale.

Despite the small numbers involved, they said Helena experience offers a clear hint that the change reduces the risk of heart attacks for smokers and nonsmokers alike from virtually the moment it goes into effect

---

C.31 hixson background

A man who snatched his 13-year-old niece after allegedly killing her parents freed the girl unharmed yesterday and surrendered after a standoff in Pennsylvania, police said.

Robert Hixson took the pajama-clad girl, Hadley Bilger, on an overnight odyssey that ended yesterday afternoon after the former Navy SEAL - surrounded by police and FBI agents - gave up in a convenience store parking lot.

At times, the 42-year-old father of three held a shotgun to his chin as he got in and out of his red pickup truck, police said. Hixson, who police said was also packing a handgun, now faces two murder charges.

The drama began Saturday night, cops said, when Hixson burst into Myron and Eileen Bilger's home in Pocono Lakes, Pa., and shot the couple dead.

As the Bilgers' other daughter, 5, ran to a relative's home to get help, Hixson grabbed Hadley and sped away.

Hadley's father, the brother of Hixson's wife, identified Hixson as the gunman shortly before he died, police said.

Police did not reveal a motive, but his mother-in-law suggested it was Hadley.

"Well, he wanted to run off with their daughter," said Donna Faye Bilger of Danville, Pa.

Before setting out for the scene of the killings, Hixson dropped his wife off at Donna Faye Bilger's house and announced he was going fishing. Then he drove more than 60 miles to the Poconos.

Pennsylvania officials issued an Amber Alert to neighboring states, notifying authorities of the child's disappearance.

About 13 hours later, at 10:40 a.m. yesterday, a tipster spotted Hixson's Chevy pickup, leading to a 45-minute chase that ended near Reading, Pa., after cops scattered tire-bursting equipment on the road.

He released Hadley, but it took four more hours for Hixson to finally surrender.

C.32 hixson new article

An ex-Navy SEAL who kidnapped his 13-year- old niece after gunning down her parents released the girl unharmed yesterday and then surrendered during a tense standoff with police , authorities

said .

Robert Lee Hixson , 42 , had threatened to kill himself with a shotgun during four hours of negotiations with police and FBI agents in **Bethel Township ,**Pa.

Hixson fatally shot **Myron Bilger Jr. , 40 , and Bilger 37-year-old wife Helen with**a shotgun after breaking into their Pocono Lake home Saturday night , police said .

Hixson , who the FBI said is an ex-Navy SEAL , is married to Myron Bilger sister .

After the shootings , Hixson abducted their daughter Hadley , leaving behind her 5-year-old sister , police said .

The sister contacted a relative , who called police .

When cops arrived , they found Myron Bilger still alive .

Before dying , he identified Hixson as the shooter , police said .

Cops would not comment on a motive , but said **they did not believe an argument preceded the shootings**.

Donna Faye Bilger , Myron Bilger mother and Hixson mother-in-law , said Hixson " wanted to run off with their daughter **. "**

Her husband , Myron Bilger Sr. , told The Post that Hixson was " a loner . "

He also said his murdered son " did n't like " Hixson .

A neighbor of Hixson in Valley Township said she heard he had moved to a trailer park there last year after getting out of jail for " hitting his ex-wife in the head with a hammer ."

Hixson , saying he was going fishing , had dropped his wife and their son at about 5:30 p.m.

Saturday at Myron and Donna Faye Bilger home in Danville **.**

When police went there after the killings , they discovered that two of Myron Bilger Sr. shotguns were missing from a locked cabinet .

Yesterday , police cornered Hixson pickup in Bethel Township , Pa. Hixson released Hadley at 11:30 a.m. and surrendered four hours later

---

C.33 jfarrell background

Colorado woman returns from Italian vacation after allegedly leaving

Police opened a child abuse investigation against a woman suspected of leaving her six children home alone while she vacationed in Italy.

Police removed the children, ages 6 to 14, from the home and began their investigation Feb. 4, the day after Jennifer Farrell left.

Police said an anonymous caller told them the children had been left alone. Police spokesman Sgt. John Gates did not return phone calls Thursday.

No one answered the phone at Farrell's house Thursday.

"Things have been blown out of proportion," Farrell told KUSA-TV in Denver after she returned from her two-and-a-half-week vacation.

Farrell's boyfriend, Hank DePetro, said the children were alone because of a breakdown in the support system Farrell put in place before she left.

DePetro said Farrell, 33, communicated with a daughter after learning by e-mail that authorities were investigating.

"I did not know what was happening until we got that e-mail and I said, 'My God, what happened?"' he said.

DePetro, 60, a retired school psychologist, is not suspected of wrongdoing, authorities said.

Weld County Social Services took custody of Farrell's son and five daughters. The younger four children are with an aunt and the two older girls are in a foster home.

Police said once they question Farrell they will turn the case over to the district attorney, who will decide whether to file charges.

Court records show Farrell has twice pleaded guilty to child abuse. The records say Farrell's former husband, Steve Farrell, had accused her of inadequately supervising the children. The couple divorced in 2000 and Jennifer Farrell had custody of the children.

C.34 jfarrell new article

SKEDADDLING MOM SKIPS TROUBLE ;

A mother who left her six children behind while she went on a two-week Italian vacation **will not face criminal charges , a prosecutor said Thursday .**

Weld County District Attorney Al Dominguez said that although he did not condone Jennifer Farrell bad judgment , he could not charge her under the child-abuse statute .

" I look at the facts .

Is there enough for me to file ? "

Dominguez told reporters .

" It not something I condone .

I do n't think it smart to leave that many children alone for two weeks ," he said .

" However , it a free country .

You have a right to make choices .

Is it a choice I would have made ?

No. "

Farrell returned to Colorado on Feb. 20 and found herself under scrutiny for having left her children with $7 , a credit card and a list of adults they could contact .

Her children range in age from 6 to 15 .

She explained that several friends had offered to look in on the children .

But because of a breakdown in communication , that did n't happen .

Instead , Greeley police took the kids into custody within 24 hours after Farrell departure , after a neighbor tipped them off .

Human Services officials placed the two oldest children in foster care .

The younger children were placed in the custody of an aunt .

For now , the children remain out of their mother custody , pending the outcome of a dependency and neglect hearing in civil court .

Farrell could not be reached for comment Thursday .

Dominguez said he had not called her to inform her of his decision .

But last week , in an interview with the Rocky Mountain News , she said essentially the same thing that Dominguez concluded .

" There no law against what I 've done ," she said .

" I 've not violated any law . "

Rocky Mountain News (Denver , CO) March 7, 2003 Friday Final Edition

In announcing his decision , Dominguez said Greeley police had determined that Farrell had developed a contact list of church members who would be available if her oldest daughter requested help .

Dominguez said there was also a main contact person who agreed to be available for the children for the two weeks that Farrell was gone .

In addition , Farrell had arranged for the children to spend a pair of three-day weekends with two different families who were members of her church .

When Farrell left on vacation , her parents were looking after the children initially , Dominguez said .

They had to leave early for medical reasons , he said .

But the grandparents had kept in touch with the children by telephone .

Several people involved in the caretaking group told investigators that the oldest daughter was mature and responsible .

The 15-year-old had attended two summers of Red Cross baby-sitting classes and infant CPR in 2001 and 2002 at Aims Community College .

Dominguez said two earlier child abuse cases against Farrell in 1991 and 1995 were not a factor in his decision .

" Maybe at some point , the legislature will make stupidity a crime , but they have n't yet ," Dominguez said .

He offered Farrell this bit of advice :

" Next time , get yourself a baby-sitter . "

NOTES : ensslinj @ RockyMountainNews.com or (303) 892-5291 GRAPHIC : Photo , Jennifer Farrell still has not regained custody of kids

---

C.35 kidnap background

Authorities on Tuesday scoffed at the claim of a 67-year-old man accused of keeping at least four women as sex slaves in a concrete dungeon that his relationship with his latest alleged victim was consensual.

Meanwhile, a scheduled competency hearing for accused serial kidnapper and rapist John T. Jamelske was postponed for a week.

"It sounds like the kind of thing that someone would say to begin building a legal defense," Onondaga County Sheriff Kevin Walsh said about the claims Jamelske, a retired handyman, made to a police officer who drove him to police headquarters on the night of his arrest on April 8.

"He knew what he was doing was a crime and he was trying to cover his bases," Walsh said. "It was all fun and games. That's his alibi."

Jamelske is charged with kidnapping, imprisoning and raping a 16-year-old girl police said he kept captive for six months in a two-room bunker he built under the yard of his suburban Syracuse home. The girl was able to escape when Jamelske took her out in public on some errands.

Since his arrest, police have confirmed that he held at least three other women captive as sex slaves. Police said one of the victims was held captive for more than two years.

A grand jury is reviewing those cases to determine what additional charges should be filed against Jamelske, who lived alone since his wife died in 1999. Police are continuing to seek out other possible victims.

In a series of"spontaneous utterances" to Manlius Police Officer B.M. Damon, Jamelske denied holding the teen captive and said he did not know she was a minor, according to a police deposition obtained by CNN.

Damon's affidavit is a sealed record, said a clerk in Onondaga County Judge Anthony Aloi's office and Damon could not be reached for comment Tuesday. Walsh confirmed that Jamelske made several statements to Damon during the ride to the police station.

Damon said in the affidavit that Jamelske spoke to him"freely and unsolicited." Jamelske said the woman"moved into his house" and that they shared many common interests, including walking his dog, karaoke singing and dancing.

"We just have fun together," Jamelske said, according to Damon's affidavit. "The only thing that she likes that I don't is blue cheese."

Jamelske said he thought the girl was 18 and that he was planning a 19th birthday party for her in May.

Defense attorney J. Michael Forsyth declined comment Tuesday about Jamelske's comments to police.

"I stand by my earlier statement that there is another side to the story that will come out at the appropriate time," said Forsyth.

Forsyth has asked Aloi to order a psychiatric exam for Jamelske to determine whether he is able to understand the charges against him and assist in his own defense. Such a request does not mean that Jamelske has mental problems, he said.

"In a case involving such bizarre allegations, I thought it would be prudent to establish one way or another his mental competency," Forsyth said.

C.36 kidnap new article

Kidnapping suspect John T. Jamelske told authorities the 16-year-old girl he is accused of holding captive in a dungeon for six months voluntarily moved into his house because they both liked walking his dog , karaoke and dancing **, a Manlius police officer confirmed Tuesday**.

" We just have fun together ," Jamelske said , according to Officer Brian Damon .

" The only thing that she likes that I do n't is blue cheese . "

Jamelske also said he thought the girl was 18 and that he was planning a 19th-birthday party for her in May, Damon said **.**

Damon confirmed thecontents of an affidavit he wrote recording Jamelske comments on the day he was questioned by authorities .

The contents had been reported in an Associated Press story .

Jamelske , 67 , of DeWitt , made unsolicited comments while Damon was driving him to the Manlius police station and to the Onondaga County Sheriff Department headquarters on April 8, after Manlius police had apprehended him with the girl , according to Sheriff Kevin Walsh **.**

Jamelske " seemed nervous , but seemed like a decent guy " during the rides , Damon said .

Jamelske , a retired handyman with a two-room concrete underground dungeon connected to the basement of his 7070 Highbridge Road home , was charged April 9 with kidnapping , rape , sodomy and sexual abuse in the case of the 16-year-old girl.

She told detectives he held her captive for six months .

Sheriff investigators have identified three other potential victims who say Jamelske held them as sex slaves in the dungeon - including one who said she was held for two years **.**

One of the victims told detectives Jamelske held her prisoner in the dungeon as far back as 1988.

" We 're still under the belief there are more victims , or people with information seeking to contact us ," said Walsh , who encouraged victims or people with information about Jamelske to call investigators at 435-3081.

Manlius police Chief Francis Marlowe and Damon , Walsh , District Attorney William Fitzpatrick and defense lawyer J. Michael Forsyth declined Tuesday to release a copy of Damon affidavit .

Forsyth said he would not vouch for the accuracy of everything that Damon alleged Jamelske said.

But he said the statements Damon attributed to Jamelske " suggest there is another side to this story. "

Walsh said Jamelske unsolicited statements suggest something else **.**

" It shows he was thinking , "

How

am I going to convince the cops that nothing wrong is going on ? "' the sheriff said .

" It shows he is competent enough to make excuses to serve his end . "

To control his victims , Jamelske threatened to kill relatives of some of his captives and made some of them write down the precise date they had sex with him , brushed their teeth or showered with a garden hose in a dungeon tub , according to Walsh .

Walsh said Jamelske had such control over the 16-year-old that she did n't even run from him when he occasionally escorted her outside .

Tuesday , Jamelske case was scheduled before Onondaga County Judge Anthony Aloi , who was to rule on a request for a formal psychiatric evaluation .

But **Aloi adjourned the case for a week because Forsyth was out of town**.

Forsyth requested the mental examination to determine whether his client is competent to stand trial and to assist in his defense **.**

Last week , the FBI sent three members of its Behavioral Science Unit from Quantico , Va. , to Jamelske house to help in the investigation .

Over two days , the agents reviewed the case compiled by Walsh detectives .

" These are people who specialize in cases that are a little more bizarre than the standard kind of case ," Walsh said .

" They have seen cases of domination and control with hostages or kidnap victims and can lend their expertise and give us their insights . "

The sheriff said he could not comment about the specific advice the FBI team gave , but he said the agents helped local investigators set priorities for a massive list of leads to check out .

" We value the additional expertise ," Walsh said

C.37 klanduke background

NEW ORLEANS –

David Duke, a one-time Ku Klux Klan leader and Louisiana politician who took his call for"white survival" overseas while the government investigated his activities, pleaded guilty Wednesday to mail fraud and tax charges.

The plea, the same day an indictment was filed, followed his lawyer's announcement Monday that Duke had returned to Louisiana to negotiate with prosecutors after three years out of the country.

Prosecutors alleged that Duke misrepresented himself to supporters as nearly broke, took in contributions and misspent the money on casino gambling and investments from 1993 to 1999. He also was accused of filing a false 1998 tax return, allegedly claiming $18,831 in income while actually having income of about $65,034.

Duke's attorney, Jim McPherson, said Monday that the Justice Department had been examining Duke for possible income tax violations involving the $100,000 sale of a list of Duke supporters to Gov. Mike Foster in 1995.

Duke had won a Louisiana House seat in 1988 and ran second for the U.S. Senate in 1990 and governor in 1991, while claiming to have jettisoned his racist views.

A poor showing in the 1992 presidential primaries effectively ended his flirtation with mainstream politics.

Duke had just started a speaking tour in Russia in January 2000 when federal agents raided his home in Mandeville, La. A search warrant, based on testimony from confidential informants, alleged that Duke took hundreds of thousands of dollars he solicited from supporters and gambled the money away at casinos.

Until his return late last week, Duke had been lecturing and speaking in Russia and other European countries in what became a crusade for"white survival" against Jews and non-Europeans.

C.38 klanduke new article

David Duke , the onetime Ku Klux Klansman and politician whose vitriolic crusade for white power and anti-Semitism seared an ill-fated trail through Louisiana government , pleaded guilty Wednesday to mail fraud and filing a false tax return .

The 52-year-old Duke has spent the last two years drumming up support in Russia for activists that he hoped would lead a worldwide wave of white supremacy .

In a plea-bargain with Louisiana prosecutors , he **faces up to 15 months in prison and $10,000 in fines for bilking supporters out of hundreds of thousands of dollars and lying about his earnings**.

For 1998 , the New Orleans native told tax collectors that he had earned less than $19,000 , according to a federal indictment .

In fact , he earned more than $65,000 that year **.**

From 1993 to 1999 , the former state senator signed and mailed pleas for help – claiming that his home and savings were about to be consumed by a pending lawsuit.

But prosecutors say Duke was n't in financial trouble .

He pocketed the checks and cash that came pouring in , or gambled the money away in **Las Vegas , Mississippi and the Bahamas, prosecutor Jim Letten said.**

Federal authorities had been investigating Duke since 1999 and negotiating with his lawyer while he traveled overseas .

Duke recently returned to the United States **to visit his ailing father in New Jersey , then arrived in New Orleans to appear before U.S. District Judge Eldon E. Fallon**.

He is free until his March sentencing **.**

Duke strange and tumultuous rise began at Louisiana State University , where he draped his room in Nazi memorabilia and gave the first of uncounted speeches about white power and Jewish conspiracy.

He eventually softened his Nazi affiliations , joined the ranks of the Klan and became grand wizard in 1975 .

Over the years , he became notorious for his outbursts **.**

He smeared his features in blackface and disrupted a 1976 Louisiana legislative ceremony honoring a black Reconstruc- tion-era governor .

He ran for the state Senate and lost.

He was linked to a riot in Jefferson Parish , and to the disruption of a civil rights march in Georgia .

In 1980 , he left the Klan and started the splinter National Assn for the Advancement of White People .

He ran for president in 1988 and lost again .

Duke finally achieved elected office in 1989 as a Republican , when he beat out a complacent incumbent to be elected state senator from the New Orleans suburb of Metairie.

As a legislative hand , he was totally ineffective – not one of his bills ever got out of committee .

The next year , Duke collected 43.5% of the vote in the primary for U.S. Senate but lost anyway .

And in 1991 , he humiliated the Republican Party and Louisiana liberals alike by beating GOP incumbent Buddy Roemer in the gubernatorial primary , but he went on to lose to Democrat Edwin W. Edwards **.**

Duke campaigns rarely turned on overt intolerance , but relied instead on the coded language that has sprung up in the New South to soften the edges of racism .

Duke railed against urban crime rates , welfare dependence and affirmative action .

" In his own way , he was effective in voicing some of the fears and alienations that many white people in the lower and middle classes feel ," said Michael Kurtz , dean of Southeastern Louisiana University graduate school .

But after his failed bid for governor , Duke faded from the state political landscape .

In 1999 , he began a tour of Eastern Europe , where he stumped for white power and worked on a pair of books – **" The Ultimate Supremacism : An Examination of the Jewish Question " and " For the Love of My People . "**

Behind the lectern , he rallied support for " racially aware " parties that he hoped would set off a " domino effect that would cascade throughout the whole world ," according to an essay posted on the Web site of the National Organization for European American Rights , a group led by Duke .

The manifesto issues somber warnings on the perils of racially mixed marriages , the " Jewish-dominated news and entertainment media " and lower white birthrates .

" Our race faces a worldwide genetic catastrophe ," Duke writes .

" Anyone who truly understands this must shoulder the immense responsibility associated with this impending catastrophe .... Our race will survive , and together we shall go to the stars!

C.39 kubby background

U.S. marijuana smoker seeks refugee status

VANCOUVER – A U.S. citizen who claims he is being persecuted in California for his use and advocacy of medicinal marijuana appeared on his own behalf yesterday at a hearing in which he is seeking refugee status in Canada.

Steven Kubby told the Canadian Immigration and Refugee Board that law enforcement officials in the state have conspired to arrest him and other sick people who have a legal right to marijuana.

Mr. Kubby, 56, a Sechelt, resident and host of a Pot-TV Web site, maintains he has to smoke large quantities to control symptoms of adrenal cancer, which generates adrenaline rushes triggering blood pressure spikes, rapid heartbeat, severe headache and chest pain.

Mr. Kubby said U.S. prosecutors are"conspiring illegally to undercut the law" – a voter-supported proposition or law that gave some Californians legal access to marijuana.

C.40 kubby new article

IN BRIEF/ SACRAMENTO ;

Canadian immigration officials put off a political refugee hearing for a month after a California medical marijuana patient fighting deportation was hospitalized over the weekend with pneumonia .

Steve Kubby , 56 , fighting adrenal cancer that he says he needs pot to control , **was suffering with a 102-degree fever and skyrocketing blood pressure , said his wife ,**Michele .

Kubby hearing began Wednesday in Vancouver .

A refugee board member delayed resumption of the hearing until April 8, but asked that Kubby provide a written update from a physician

---

C.41 leung background

A U.S. magistrate judge refused today to release alleged Chinese spy Katrina M. Leung on bail, saying he is not certain investigators had found all of the classified documents she may have procured during two decades of alleged spying.

U.S. Magistrate Judge Victor Kenton also said federal prosecutors had convinced him during a four-hour hearing here that a significant portion of Leung's financial assets were in Hong Kong, making it easy for her to flee.

Federal prosecutors allege that Leung, 49, who was paid $1.7 million as an FBI informant for 20 years, was instead a"double agent" who pilfered classified material from her handler, former FBI agent James J. "JJ" Smith, and gave it to the People's Republic of China. Smith, 59, has been charged with gross negligence for allowing her access to the material.

The case is complicated by a 20-year sexual affair between Leung and Smith, both of whom are married to others, court documents show.

A second former FBI Chinese counterintelligence official also had an intermittent affair with Leung that ended in 1999, court documents show. Authorities have identified the second former agent as William Cleveland Jr., who resigned his counterintelligence post at Lawrence Livermore National Laboratory last week after being stripped of his security clearances.

Government officials argue that Smith and Cleveland failed to properly alert senior FBI officials to suspicions about Leung, and continued to have romantic relationships with her. Cleveland in early 1991 discovered an audio intercept indicating that Leung had tipped off the Chinese to a diplomatic security trip, and he warned Smith about her duplicity, according to records and sources.

But attorneys for Leung, who is well known in California Republican political circles, argued in court today that Smith and others at the FBI encouraged her to provide information to Chinese intelligence to win their trust. "There's nothing to suggest that she passed anything to any body without the consent of this FBI," said Leung's attorney, Janet I. Levine. "The only documents she had access to were brought to her."

Smith's defense team, meanwhile, argues that senior FBI counterintelligence officials in Washington were long aware of problems with Leung. Smith is free on $250,000 bond.

According to court documents and law enforcement officials, both Cleveland and Smith attended a May 1991 meeting at FBI headquarters in Washington that included discussion of Leung's unauthorized contacts with agents of the Chinese Ministry of State Security, according to sources familiar with the meeting.

"The discussion was, 'We think we may have some problems with a source who's handing back information to the Chinese, but JJ Smith is handling it,' " one official said. "Nobody ever followed up on it."

Smith's attorney, Brian Sun of Los Angeles, said the FBI"is finally acknowledging what I said early on: that senior FBI officials were fully aware of the potential issues with Leung. To try and hang it all on my client would not be accurate."

FBI officials have told lawmakers that the Leung case means that every major Chinese counterintelligence case over the last 20 years is potentially compromised, including those involving espionage, technology theft and alleged efforts by the Chinese government to influence the 1996 presidential elections.

Eggen reported from Washington.

C.42 leung new article

Calling her a flight risk and potential threat to national security, a federal magistrate denied bail Tuesday to Katrina Leung , a longtime FBI informer accused of working as a Chinese double agent .

U.S. Magistrate Victor Kenton ordered the **wealthy San Marino businesswoman**jailed at the **federal Metropolitan Detention Center**pending her trial on a charge of illegally obtaining classified documents from her FBI handler , with whom she carried on a 20-year sexual relationship .

The agent , James J. Smith , now retired , was charged with gross negligence in the handling of national security documents and freed on $250,000 bond following his **arrest April 9 along with Leung**.

During a four-hour hearing before Kenton on Tuesday , Leung lawyers argued that she was more deserving of bail than Smith .

At one point , defense attorney Janet Levine **wondered aloud whether prosecutors were motivated by prejudice in seeking detention for Leung , a naturalized U.S. citizen who was born in China .**

Assistant U.S. Atty .

Rebecca Lonergan denounced the suggestion as outrageous and untrue.

Lonergan contended that if allowed to go free on bond , Leung **might flee to China , where she has friends in high places and which does not have an extradition treaty with the United States .**

Although Leung , 49 , and her family had offered to post **as much as $2 million in property**to secure her freedom , Lonergan argued that the defendant had possibly **millions of dollars**stashed away in hidden foreign bank accounts , money she could use if she chose to flee .

The prosecutor **described Leung foreign assets as " enormous and complex ," and charged that the suspect had concealed her overseas earnings in her annual U.S. tax returns .**

Lonergan also asserted that while working as a paid informant **, Leung made about 15 overseas trips between 1989 and 2002 without telling her FBI handler .**

And just last month , the prosecutor said , Leung was offered a five-year visa to China during a meeting with the deputy chief of mission at the Chinese Embassy in Washington .

But defense attorney Levine and co-counsel John Vandevelde **accused the government of distorting the facts about Leung conduct .**

Levine said that Leung has had opportunities to flee since December,when FBI agents searched her home and interviewed her at length over seven days.

During her interrogation , Vandevelde said **, Leung cooperated with the FBI , volunteering information and documents , despite the fact that she had been presented with a search warrant that said she was suspected of committing an act of espionage carrying a possible death sentence .**

The actual charge against her carries a maximum 10-year prison term .

Vandevelde also said that **Leung gave advance notice to FBI investigators about her meeting with the Chinese diplomat in Washington last month and reported afterward on what transpired.**

" Ms. Leung was exceedingly candid and forthcoming with the FBI ," said Vandevelde.

" In contrast , agent Smith withheld information when he was questioned in this investigation . "

The defense lawyers suggested that Leung be placed on home detention with electronic monitoring of her movements.

Leung husband , brother and sister attended the hearing and indicated they were willing to post assets for her bail .

In deciding against releasing her , Kenton cited an FBI statement that the agency has been forced to review a number of national security cases to determine whether they were compromised by unauthorized information that Leung might have passed to China **.**

" The court cannot conclude that the defendant does not pose a danger to national security ," he said

---

C.43 makemeth background

Desperate to halt a soaring drug problem in rural Missouri, state lawmakers are weighing severe restrictions on sales of common over-the-counter cold medications, such as Sudafed, that can be used to make methamphetamine.

The House last week passed the toughest legislation in the nation regulating pseudoephedrine, the active ingredient in most nasal decongestants. The bill, now under consideration by the Senate, would limit customers to two boxes of medication per transaction.

More controversial still, the bill would require stores to keep nasal decongestants behind the counter or within 6 feet of the cash register, or to tag each box with an anti-theft device.

Several other states have banned consumers from buying more than three boxes of cold pills at one time. But no other state regulates where decongestants can be sold, said Nancy Bukar, a lobbyist for the Consumer Healthcare Products Assn., which represents manufacturers and distributors of over-the-counter medicines.

"This is a ridiculous solution," said Ronald Leone, executive vice president of a trade group representing convenience stores. Most stores, he said, already stash cigarettes, adult magazines, liquor and condoms behind the counter. There's no room, he argued, for decongestants. "This law is draconian," he said, "and it's not going to solve the problem."

The problem is that pseudoephedrine can be combined with other ingredients – such as anhydrous ammonia from farm fertilizer or red phosphorus from matches – to produce methamphetamine, a highly addictive, illegal stimulant that is also known as meth, ice, crystal and crank.

Missouri law enforcement officers raided 2,725 meth labs last year – an average of more than seven a day. California's illicit labs tend to produce a higher volume of the drug, but for two years running, Missouri has led the nation in the sheer number of meth seizures.

The makeshift labs – which involve highly volatile chemicals and are prone to explosions – have been found all over the state, including wealthy suburbs. They are most common in sparsely populated rural areas, where meth addicts convert abandoned barns into miniature factories, using everyday products such as cold pills, propane tanks and coffee filters to brew the stimulant.

"Methamphetamine represents the fastest-growing drug threat in Missouri," said Sen. Anita Yeckel, a Republican who is sponsoring the pseudoephedrine bill. "We want to take the most aggressive way we can to fight the problem.... Tighter control of meth ingredients seems to be one of the most promising approaches."

Critics argue that although the bill might slow down theft of cold pills, it would not prevent a meth addict from buying as much as he needed to make the drug. Although customers would be limited to two boxes per transaction, they could get back in line again and again, buying two additional boxes every time. Or they could buy two at each of a dozen stores.

Still, Capt. Chris Ricks of the Missouri State Highway Patrol said law enforcement would welcome any restrictions that make it even a bit harder for meth cooks to get their hands on pseudoephedrine. "There is a methamphetamine epidemic in this state," Ricks said. "We're looking for any tool we can get that might help us fight it."

C.44 makemeth new article

JEFFERSON CITY , Mo. –

Pharmacist Greg Mitchell knows the routine .

Customers come into his pharmacy asking for several boxes of Sudafed – not even caring to hear about its proper use or showing signs of stuffed-up noses **.**

" They just want to buy the stuff in quantities and go ," says Mitchell , who now refuses to sell more than one package at a timefrom his pharmacy in Lexington , 35 miles east of Kansas City .

The medicine can be used by makers of the illegal drug methamphetamine .

Responding to similar scenarios around the state , Missouri lawmakers are proposing some of the nation toughest restrictions on the sale of over-the-counter medicines such as Sudafed .

Their motivation : Missouri police seized a nation-high 2,725 clandestine meth labs last year – nearly one out of every five labs found nationwide **.**

Pseudoephedrine , the sole active ingredient in decongestants such as Sudafed, also is a key ingredient in methamphetamine , a powerful and highly addictive stimulant that law enforcement officers label a huge problem in the Midwest , Southwest and West .

The Missouri legislation , which already has passed the House and awaits Senate debate , would require medicines such Sudafed to be placed either behind the counter or within six feet of a cashier or to contain an electronic anti-theft tag .

It also would limit each customer to two **packages , or 6 grams**, of pseudoephedrine medicines .

Missouri currently is **one of just six states**that now impose a **three-package limit on**such medicines **.**

The North Dakota Legislature passed a bill last month imposing a two-package limit and prohibiting sales to anyone younger than 18 .

An Oklahoma law was passed last year in an effort to put criminally ambitious retailers in jail alongside their customers .

The law makes it a crime for someone to knowingly to sell precursors to methamphetamine manufacturers **.**

Offenders can face a 10-year prison term .

Law enforcement officers said the Oklahoma law also makes it illegal for someone besides retailers , pharmacies , drug manufacturers or health-care professionals to posses more than 24 grams of ephedrine , pseudoephedrine or phenylpropanolamine .

Anyone who has more than that amount is presumed to possess it with the intent to manufacture an illegal drug and could receive a five-year prison term.

In another , apparently unprecedented , move **, Oklahoma filed a civil lawsuit in October against six of the state major suppliers of pseudoephedrine .**

" This is the first state action of its kind in Oklahoma and perhaps the nation ," state Attorney General Drew Edmondson said then .

The lawsuit claimed that the companies negligently sold large amounts of pseudoephedrine , often under suspicious or illegal circumstances, and that the narcotics ultimately ended up in illegal methamphetamine labs **.**

Most of the companies ' sales were to convenience stories and gas stations , " which have been identified by DEA as primary sources of illegally diverted pseudoephedrine ," the lawsuit said .

Although law enforcement agencies in Oklahoma and elsewhere have persuaded many retailers to restrict voluntarily the sale of products that can be used to make methamphetamine , no state presently imposes the type of retail display restrictions being considered in Missouri , according to the Consumer Healthcare Products Association .

The group represents manufacturers and distributors of over-the-counter medicines **.**

" This proposal just goes a little too far ," said Mike Sargent , a lobbyist for the association .

" It really unfairly targets the chronic allergy sufferer , because that the consumer who uses this product most often . "

Convenience stores also have objected to the Missouri proposal , saying it would force decongestants to be moved to prime counter spots that now are saved for impulse items such as candy and gum.

The alternative would be to put cold and allergy medicines alongside tobacco products behind the counter – **making it difficult for shoppers to compare products , said Ronald Leone , executive vice president of the Missouri Petroleum Marketers and Convenience Store Association .**

" All it really going to accomplish is overregulating businesses and keeping lawful products out of the hands of lawful consumers ," Leone said .

But some stores and city governments already have taken the precautions proposed in the legislation.

In St. Peters , Mo. , the City Council enacted an ordinance last fall requiring pseudoephedrine medicines to be placed behind the counter or in locked display cases

.

The proposed state law would override the city more restrictive ordinance , as well as similar local laws adopted in other Missouri cities .

State legislators " are making a sham out of this ," said St. Peters Mayor Tom Brown.

" They 're completely undermining the protection we sought for the residents of our communities and the citizens who work in retail stores and sell these products . "

The St. Peters ordinance was enacted partly because of the May death of grocery store security guard James Toppett, who was smashed between a wall and a pickup while chasing two people who were suspected of stealing several packages of decongestants.

Police say the suspects were using the drug for methamphetamine **.**

Since the local law took effect Dec. 1, no thefts of Sudafed or other pseudoephedrine drugs have been reported , said St. Peters Police Capt. Jeff Finkelstein , who helped craft the ordinance .

State Rep. Rob Mayer, a Republican from rural southeast Missouri who is sponsoring the legislation , points to places such as St. Peters as the impetus for his idea **.**

He also cites national crime statistics , which show that Missouri methamphetamine lab seizures – already the highest in the nation in 2001 – rose an additional 29 percent in 2002 .

Retailers around the state need a uniform law , he says.

And in most places , his proposal would be tougher than local ordinances .

The meth problem began several decades ago in California , which still has some of the largest producers and the second-highest number of lab seizures .

It spread east during the past decade or so and has taken root especially in the Midwest , where rural areas provide cover for small , makeshift labs **that often produce a stinky , rotten-egg smell .**

" Methamphetamine is the No. 1 drug threat to rural America ," said Rusty Payne of the federal Drug Enforcement Administration, which on **Tuesday announced the arrests of more than 65 people in the United States and Canada for illegally importing pseudoephedrine .**

Mitchell , the pharmacist , worries that the Missouri legislation is too sweeping , removing from easy consumer access a product that is used legally by most purchasers .

But he open to the possibility of more restrictions on his pseudoephedrine sales.

" But the bottom line is to cut into this manufacture of meth , and reasonable steps would be OK ," Mitchell said

---

C.45 medicaid background

Pharmaceutical giants Bayer and GlaxoSmithKline on Wednesday reached a record-setting Medicare fraud settlement over allegations they overcharged millions of dollars for popular prescription drugs.

Bayer, based in Germany, agreed to pay $257 million to settle allegations that it overcharged Medicare for prescription drugs including Cipro, the popular antibiotic used during the U.S. anthrax scare. A Bayer spokesman said the company did not believe it had done anything illegal. The company set aside $257 million last year to settle the case.

Glaxo, based in Britain, agreed to settle the case for $87.6 million"to avoid delay and expense of a trial," the company said. Glaxo said it still believes it was acting in good faith. The government said that Glaxo overcharged for the antidepressant drug Paxil and Flonase, a nasal spray for allergies.

"These frauds impacted the most vulnerable citizens we have, the elderly and the poor," said Assistant U.S. Atty. Susan Winkler.

The settlement money will be divided among the federal government, states and some public health entities. The states will receive $147.3 million. The California attorney general's office said Wednesday that the state's share could amount to $32.2 million.

The investigation focused on allegations that the two companies hid their lowest prices from Medicaid by repackaging or relabeling their drugs under a middleman's name. The middleman then sold the drug at a deep discount not reported to the government.

By law, the companies are required to pay Medicaid a rebate if they charge anyone less than the government.

The case was initiated by a Bayer executive who sued under the federal False Claims Act, which allows citizens to sue contractors on behalf of the government.

As part of the settlement, heirs of the now deceased whistle-blower, former Bayer marketing executive George Couto, will receive $34 million, according to the family's attorney.

"The government is sending a message that defrauding and abusing government health programs and their beneficiaries simply will not be tolerated," said Amy Wilken, associate director of the Taxpayers Against Fraud-the False Claims Act Legal Center in Washington.

Times wire services were used in compiling this report.

C.46 medicaid new article

Bayer AG **, Germany biggest drug maker**, and U.K. rival GlaxoSmithKline PLC settled allegations that they overcharged the U.S. government Medicaid health plan for the poor .

Both companies had set aside money for the case .

Bayer , whose U.S. **operations are based in Robinson**, will pay $257 million , the company said in a statement yesterday , noting that it had set aside that amount **in December and**took a charge for it in the fourth quarter .

Glaxo will pay $87.6 million .

The United States is investigating drug makers ' compliance with rules that require them to give the government the same deals as private customers .

The **four-year probe has**targeted several drug **makers including Pfizer Inc. and Schering-Plough Corp.**Prosecutors said Bayer and Glaxo repackaged drugs and sold them under the label of a middleman to hide their lowest prices from the government .

Drug companies are required to give their lowest price to Medicaid and to pay rebates if they charge any other customers less .

Bayer was accused of overcharging the U.S. government for its antibiotic Cipro , used to treat infections including anthrax , and its **high blood pressure drug Adalat .**

Glaxo was accused of overcharging for Paxil , an antidepressant , and Flonase , an allergy spray .

" This settlement resolves a long-standing investigation of the company marketing practices , which Bayer believed were responsible and conducted in good faith ," Bayer said in its statement .

" We felt our interpretation was reasonable and in good faith ," Glaxo spokeswoman Mary Anne Rhyne said yesterday .

" We received a clarification and we 're living by that . "

The payments will be made under an agreement with the U.S. attorney office in Boston.

The settlements , which include 49 states and the District of Columbia , represent the largest national Medicaid fraud settlements ever , Pennsylvania Attorney General Mike Fisher said .

Pennsylvania share of the settlements is about $11 million , which will be returned to the state Medicaid program , he said .

George Couto **was a senior marketing executive for Bayer in Connecticut in February 2000**when he contacted the Justice Department , triggering the investigation .

He accused the company of knowingly providing incorrect data on the prescription drug prices and preventing the government from receiving discounts .

Couto **, who died of cancer in November,**claimed that Bayer cheated the Medicaid program out of more than $100 million in rebates , according to his lawyer , Neil Getnick .

Couto sued under the federal False Claims Act , a law that allows citizens to sue contractors on behalf of the government .

His heirs will receive $34 million from the Bayer settlement , according to Getnick .

NOTES : Bloomberg News and Post-Gazette staff writer Patricia Sabatini contributed to this report

---

C.47 molest background

A man whose private, fictitious account of sexually torturing children landed him in prison is finding support among advocates of the First Amendment, which protects free speech.

Brian Dalton, who pleaded guilty to pandering ob-scenity involving a minor rather than go on trial and have his journal made public – or risk a more severe sentence – decided last week to try to withdraw that plea.

If successful, the 22-year-old Columbus man could find himself the banner carrier for civil libertarians outraged by an Ohio law that could punish a person for merely writing his or her thoughts.

And he could be characterized by supporters of the prosecution as a potential pedo-phile who must be stopped.

Dalton was sentenced to seven years in prison July 3 when he pleaded guilty in Franklin County Common Pleas Court to a pandering charge.

Typically, pandering charges in-volve photographs or images of children engaging in sexual activity. But the state law was applied to Dalton for something he wrote.

On the day of the conviction, Franklin County Prosecutor Ron O'Brien called the case a breakthrough in the application of child- pornography laws, but he has since remained silent on the subject. He has instructed assistant prosecutors not to discuss the case.

But that hasn't stopped others from weighing in.

The New York Times and Washington Post have reported on the case. Major television networks in the United States and England also have covered the story.

Many question the constitutionality of the Ohio law that O'Brien's staff used to prosecute Dalton.

Critics challenge the prosecutor's office for bringing the charge, saying Dalton was prosecuted and convicted for thinking bad thoughts, not performing bad acts.

Others criticize Dalton's attorney for allowing Dalton to accept the plea bargain.

Dalton on Thursday asked to withdraw his guilty plea. And the American Civil Liberties Union of Ohio is considering assisting Dalton's attorney in the battle. "Brian did not realize it was going to take off like this," said Isabella Dixon, Dalton's attorney. "However, the spotlight is not something he's looking for."

Dalton, who would not talk about the case, is concerned about the journal becoming public, something that would happen if the case went to trial, Dixon said.

"He's thinking in terms of what happens when more of the facts come out," she said. "He is well aware (that) once this journal is made public, that things other than the pornographic things would be offensive to people."

Dalton admits writing the journal, in which he describes killing two crackheads and taking their children home, where he places them in a cage in the basement. He also describes in detail binding, torturing and sexually assaulting a young boy and girl.

"But with all the media attention, it has caused him to really, really look at the constitutional issues," Dixon said. "Before, he was focused on the trial."

Dixon said she apprised Dalton of his options, but it was his decision to plead.

She believes Dalton would have been convicted had he gone to trial.

"I think once 12 people read the contents of that journal, they were probably going to find him guilty of the law the way it is written," Dixon said. "Once he appealed, it would be a different situation."

Dalton was on probation for a 1998 conviction for possessing pornographic photographs of children when his probation officer found the 14-page journal during a routine visit of his home earlier this year.

The day of the plea, Assistant Prosecutor Christian Domis said there were concerns that the case infringed on First Amendment rights, but he believed Dalton's next step would be to act out his fantasies.

Domis also believed Dalton's case fit the law.

The Feb. 23 indictment alleges that Dalton"did create, reproduce or publish any obscene material that has a minor as one of its participants or portrayed observers."

The charge was filed because Dalton"created" the journal, Domis said the day of the plea.

Court records show that this was not the first time Dalton had written about sexual fantasies involving children.

In the 1998 case, Dalton was charged with 20 counts of pandering obscenity and sexually oriented matter involving a minor for possessing pictures of girls who appeared to be between 7 and 15 years old engaged in various sexual acts.

The child pornography was discovered in plain view by maintenance workers who were in Dalton's apartment to work on the heating and cooling system.

The prosecutor in the case said police were alerted and obtained a search warrant. They not only uncovered the pornographic pictures, but also a" handwritten story describing the brutal rape and torture of a little girl," according to the court transcript.

"He acknowledged writing the story and described it as a fantasy. A small girl described in the story was Mr. Dalton's 10-year-old cousin," the transcript said.

However, Dixon and prosecutors say that the 1998 charges involved only the pornographic pictures Dalton had downloaded from the Internet, not the writing.

First Amendment experts say Dalton should not have been charged for writing or owning the journal recovered earlier this year.

"It is possible to make an argument that his probation should have been revoked, but his prosecution for a separate offense, you can't do it," said David A. Goldberger, a law professor at Ohio State University who specializes in First Amendment law.

The Ohio law raises questions concerning invasion of privacy and criminalizing thought processes when, in fact, no unlawful conduct has occurred, he said.

Others see it differently.

Bruce Taylor, president and chief legal counsel for the National Law Center for Children and Families, a nonprofit educational and legal advice organization in Fairfax, Va., said he has been a"spokesman in a sense" for the Ohio law in the Dalton case, which he supports.

"It's all over this country – everybody wants to talk about it," Taylor said.

Taylor, who prosecuted many obscenity cases – including one involving Larry Flynt of Hustler magazine – as an assistant city prosecutor in Cleveland, said Ohio's law is probably unique in criminalizing the creation of child pornography.

"If Dalton had been caught distributing it, then that would have been a crime under adult obscenity laws, regardless of whether it was child pornography or not," Taylor said.

"But just finding it in his possession meant it was probably only prosecutable under Ohio law because it (the law) includes not only possession of images, but possession of written obscene child pornography.

"I think Ohio law probably is unique in that respect. That's why it is the first case of its kind."

Benson Wolman, a First Amendment lawyer and former executive director of the ACLU of Ohio, says other states may have provisions similar to Ohio's, but other prosecutors have used more restraint.

"I think some prosecutors may look at it and see the obvious unconstitutionality and, as a result, they choose not to bring actions similar to this," Wolman said.

In 1990, O'Brien argued an Ohio case to the U.S. Supreme Court that upheld making the private possession of child pornography a crime.

"It used to have to be real children, and that's the reason you criminalized the possession or at least the distribution of child pornography – because real children got hurt," Taylor said.

The Ohio law doesn't punish people for their fantasies, he said, but for writing them in the form of child pornography.

"This statute told Brian Dalton and everybody else, 'I can't stop you from having bad thoughts, but I'm going to prohibit you by statute from writing them down, taking pictures of them and reducing them to a physical form of an obscene form of child pornography,' " Taylor said.

"Because then you've created a dangerous instrumentality that is eventually going to lead to some real children getting hurt in the future, either because of the effect on the creator or other pedophiles who get their hands on it in the future."

Taylor said the law has narrow application directly to child pornography.

"It has to be obscene, which means it has to be explicit. It has to appeal to the prurient interests of pedophiles, not normal people. It has to lack serious literary, political or scientific value."

Wolman said there are exceptions to the law – but not an explicit exception for artistic or literary work.

"Nor does it say anything about a personal journal," he said.

Wolman testified before a legislative committee concerning the statute in the late 1970s and again in the 1980s.

"I'd be surprised if legislators really thought that if somebody wrote something down in their personal journal that that would be a felony."

Currently, a U.S. Supreme Court precedent says child pornography requires a visual depiction – a photograph – of a child involved in sexual conduct.

"But laws get expanded by challenges such as this," said Janet LaRue, senior legal studies director of the Family Research Council, a nonprofit, public-policy organization in Washington, D.C.

LaRue supports expanding the law to protect children but isn't confident the Dalton case would withstand a legal challenge.

"Obscenity can be words, but it has to be distributed. Possessing it privately, especially when you created it yourself, isn't criminal," she said. "With child pornography, possession is already illegal, but it has to have a photograph under current law."

C.48 molest new article

Columbus -

A legal posse of Ohio top civil libertarians yesterday rode to the defense of a Columbus man jailed for private diary entries that described sexual fantasies involving children.

The team - led by American Civil Liberties Union-Ohio founder Benson Wolman , Ohio State University law Professor David Goldberger , and **Cleveland-based ACLU attorneys -**intends to fight Brian Dalton conviction on the grounds that his private account of caging and sexually torturing two young children is protected by the First Amendment .

Wolman and Goldberger are no strangers to controversy on behalf of the Constitution **.**

In the 1970s , Goldberger handled the ACLU defense of neo-Nazis who wanted to demonstrate in predominantly Jewish Skokie , Ill.

In the 1990s , Wolman , who is Jewish , defended the Ku Klux Klan right to place a cross on the grounds of the Ohio Statehouse next to a menorah and a Christmas tree .

Civil libertarians across the country , concerned that First Amendment rights are being chipped away , hope to capitalize on the attention that Dalton case has attracted to make a nationwide splash .

A Franklin County Common Pleas judge on Tuesday rejected Dalton request to withdraw his guilty plea to a charge of pandering.

In exchange for the July 3 plea , a second pandering charge was dismissed .

After his case got national attention , Dalton asked last month to withdraw the plea .

But **Judge Nodine Miller this**week **said that Dalton had made his plea knowingly and that a claim that he had been promised treatment instead of prison time was unfounded .**

She added that Dalton attorney , Isabella Dixon , never raised constitutional questions about Ohio pandering law during the case .

Wolman **, who has read the journal, agrees that " the vast majority of reasonable people would agree it is disturbing," but said no one should be in jail for " things he scribbled in his notebook ."**

Dalton , 22 , has said that the writings were fictional and never distributed to anyone **, causing free-speech advocates nationwide to attack Ohio pandering statute for jailing people for " thought crimes . "**

Dalton parents found his journal in their son apartment and turned it over to authorities monitoring his probation on an earlier child pornography offense.

Dalton parents said they hoped that by turning over the journal that their son would receive treatment to overcome his sexual attraction to children , not a second criminal conviction and seven more years in prison .

He was convicted in 1999for possessing child pornography and **was paroled after four months.**

Wolman said Dalton was returned to Madison Correctional Institution in July for violating the conditions of his parole and has not yet begun his sentence on the pandering charge .

Wolman said the fact that constitutional issues " never surfaced " in either of Dalton previous cases could help lawyers who continue to fight for his freedom .

" It may be necessary to assert ineffectiveness of counsel ," Wolman said .

" That not to say he had incompetent counsel , because all lawyers make mistakes .

But I do n't think grasped the enormity of this issue . "

Dixon did not return calls yesterday .

Wolman said there are several other legal options being considered , including asking the judge to reconsider or appealing her decision .

Contact Julie Carr Smyth at

C.49 nukeleak background

Engineers at a nuclear plant near Bay City are examining residue, about half the size of an aspirin, which was found last Saturday while the one of the plant's reactors was shut down for scheduled refueling and maintenance, representatives for the South Texas Project Electric Station said Friday.

The powdery material was found on the outside of two instrument guide tubes where the tubes enter the bottom of the reactor, the plant representatives said.

The reactor is encased in a concrete and steel-lined containment building.

Test results indicate the residue came from reactor coolant fluid, plant officials said.

"Finding this minute amount of residue demonstrates our inspection process works," said Ed Halpin, the plant's general manager.

The plant said it is working with the Nuclear Regulatory Commission to review the needed corrective actions. The plant said a team of engineers and chemists have reviewed all of the instrument guide tubes and found no additional residue.

"Our priorities are to determine the cause of the problem and make sure our corrective actions are effective," Halpin said, adding the unit will remain out of service until the problem is fixed.

The plant's other unit remains online and is operating at full power. The plant's two reactors combine to produce more than 2,500 megawatts of electricity.

The plant supplies power to customers from Houston to Austin and San Antonio to Corpus Christi.

C.50 nukeleak new article

WADSWORTH –

Engineers at the nuclear plant near Bay City are examining residue about half the size of an aspirin that was found **April 12 while**one of the plant reactors was shut down for refueling and maintenance , representatives for the South Texas Project Electric Generating Station said Friday .

The powdery material was found on the outside of two instrument guide tubes where the tubes enter the bottom of the reactor , the plant representatives said .

The reactor , which is partly **owned by the City of Austin**, is encased in a concrete and steel-lined containment building .

Tests indicate the residue came from reactor coolant fluid , plant officials said .

" Finding this minute amount of residue demonstrates our inspection process works ," said Ed Halpin , the plant general manager **.**

Officials with Austin Energy said the reactor problem wo n't affect their delivery of electricity .

But if the unit remains shut down for several months , it could mean higher electric bills for customers .

Because the downed unit is taking 200 megawatts away from Austin Energy , the utility will have to make up for it by burning costlier natural gas , said Ed Clark , the utility spokesman.

Consumer bills will be noticeably higher only if the loss continues into the hottest part of the summer.

One megawatt is enough to power 200 homes .

The plant said it is working with the Nuclear Regulatory Commission to review the needed corrective actions .

The plant said a team of engineers and chemists has reviewed all of the instrument guide tubes and found no additional residue .

" Our priorities are to determine the cause of the problem and make sure our corrective actions are effective ," Halpin said , adding that the unit will remain out of service until the problem is fixed .

The plant other unit is operating at full power .

The two reactors combine to produce more than 2,500 megawatts for customers from Houston to Austin and San Antonio to Corpus Christi

---

C.51 ohiostate background

Columbus - An Ohio State University sophomore and four others who attended his 21st birthday party died yesterday in a fire that destroyed an off-campus house.

Two male students from OSU and three women from the Alpha Gamma Delta sorority at Ohio University in Athens were killed, police said.

At OSU, students openly wept outside the burned building throughout much of the day, mourning the loss of their young friends and questioning what had caused the fire to ignite, then spread so quickly.

Amanda Shak, a sophomore from Lima, described the fire as"something so horrible, you can't imagine" and said flames shot from the doors and windows of the three-story brick house, just east of the campus.

At OU, a sign outside the Alpha Gamma Delta sorority house read, "House closed - friends and family only." Dozens of students gathered outside the large brick sorority house, hugging and sobbing. A nearby fraternity brought red, white and yellow roses to an impromptu memorial service.

The cause of the fire is under investigation, and police would not release the names of the deceased until the coroner's office confirmed their identities by checking dental records. However, it was reported that the party was celebrating the 21st birthday of Alan R. Schlessman, a business major from the Sandusky area. Schlessman's grandfather, Richard Widdoes, last night, said that Alan was one of the fire's victims.

Witnesses said that the fire occurred after a birthday celebration - an event they described as largely upbeat, although it was briefly interrupted by a shoving match between two unidentified males.

Police said they are trying to determine whether the apparent underage drinking that occurred at the party played any role in the fire.

Homicide detectives and arson investigators are in charge of the probe, said Columbus police Sgt. Brent Mull. Although he called the fire"suspicious," he said arson investigators routinely are called in when a fire results in death.

"This fire is one of the most deadly and tragic to ever have occurred in the city of Columbus," Mayor Michael Coleman said in a news conference.

Columbus police could not recall a deadlier fire in decades, and said it could be the city's worst blaze since a fire at the old Ohio Penitentiary killed more than 300 people in the 1930s.

About 100 students attended the party on East 17th Avenue, many of them sophomores who have not reached the legal drinking age of 21, according to several attendees who declined to give their names.

Most guests left by 3 a.m., witnesses said, although more than a dozen people were still in the house - some sleeping - when firefighters arrived.

Fire officials were called about 4:05 a.m. When they arrived, they found the house engulfed in flames. After rescuing OSU sophomore Josh Patterson of Cincinnati and two others, they extinguished the fire at 5:24 a.m., said Battalion Chief Mike Fultz.

Patterson, 20, was in critical condition and being treated for minor burns and smoke inhalation at Ohio State University Medical Center, a hospital spokesman said.

Jennifer Lehren, 20, an OU student from Centerville, was treated and released.

After returning to the house with her hands bandaged and hair singed, Lehren said that she and Patterson - her boyfriend - were asleep in his bedroom at the back of the house on the second floor when she woke up while being carried out.

"I was incoherent. I didn't know what was going on until I was outside," she told the Associated Press. "I remember screaming that it was so hot and that my hair was on fire."

OSU President Karen Holbrook commended firefighters for their bravery and quick response, and credited them with"saving the lives of several who were in the house."

Two firefighters also were treated for minor injuries and released, Fultz said.

Stephanie Weaver, a sophomore from Lima who lives three doors down from the blaze, said she attended the party and was sleeping when the fire began.

"The police came and knocked on the doors and told everybody to get out," she said.

Weaver and other evacuated neighbors stood across the street and watched as firefighters searched the house and battled the blaze, which eventually spread to houses on either side.

As she and others discussed the fire, Mayor Coleman walked up to offer words of comfort.

"Did you know them?" Coleman asked. Weaver and her roommate, eyes filled with tears, said they did. "I'm sorry. I'm so sorry," the mayor said, then moved on to the next group.

Police estimated that the house where the blaze began sustained about $250,000 in damage. The houses on either side were not damaged as severely, although cost estimates were not available.

Chief Fultz said witnesses reported hearing smoke alarms go off in the house although he said firefighters did not hear the alarms. The agency is investigating whether the alarms were working properly.

Plain Dealer Reporter Mike Tobin and Ohio University journalism student Anthony Castrovince contributed to this story.

To reach this Plain Dealer reporter:

C.52 ohiostate new article

The fire that killed five college students in **an apartment near**Ohio State University **has led to what Columbus police say is the largest murder investigation in the city history .**

" We have decided this is an incendiary arson fire that started near the front of the building ," Columbus Fire Capt. Steve Saltsman said during a news conference yesterday .

He said evidence shows the fire at **64 E. 17th**Ave. **was set about 4 a.m.**

Sunday after a party .

He and other officials refused to provide details about the evidence , except that samples from the front of the house were being tested at the state fire marshal forensics lab .

The case is the largest murder investigation in Columbus considering the number of victims , homicide Lt. Mary Kerins said.

Wendell Moore fatally shot four people , including his wife , daughter and niece , in May 1987 .

Jerry Hessler killed three people in Columbus and a fourth in Worthington on the evening of Nov. 19, 1995 .

Columbus police helped investigate the slayings of 10 people by the Lewingdon brothers in 1977-78 , but only three of the victims were killed in the city .

Kerins ' detectives yesterday asked for the public help in identifying a suspect in the 17th Avenue fire .

A reward that yesterday totaled at least $20,000 is being offered for information leading to the arrest and conviction of a suspect **or suspects .**

Anyone with information is asked to call police at 614-645-4730 or Crime Stoppers at 614-645-TIPS .

" We need the public help .

We 're looking for anyone with videotapes of the fire ," Kerins said .

The fire claimed the lives of Ohio University students **Christine Wilson , 19 , of Dublin ; Andrea Dennis , 20 , of Madeira ; and Erin DeMarco , 19 , of Canton .**

Also killed were Ohio State students Alan Schlessman , 21 , of Sandusky , and Kyle Raulin , 21 , of West Chester .

Franklin County Prosecutor Ron O'Brien said a suspect would face charges ranging from involuntary manslaughter to murder.

The five students had gathered Saturday night with about 80 others at **the three-story rooming house**to celebrate Schlessman 21st birthday **with four or five kegs of beer .**

When the fire started , about 20 people were left inside .

On a recording of Fire Division radio traffic that morning , a firefighter can be heard saying , " There was a lot of intoxicated people in there . "

After the news conference yesterday , Bill Hall , vice president of student affairs at OSU , said the fire could serve as a " rude awakening " to some.

" We 've tried to say that with these large parties , you lose the ability to control the situation ," he said .

Some inside the house reported hearing a pop or breaking glass when the fire started , but officials declined to comment on that yesterday .

At the news conference , police said they have no suspects and do n't know of a motive for the fire .

They discounted the idea that a fight Saturday night in the house might have led to it .

" To characterize it as a fight would be an injustice ," said detective Mike McCann .

" It was a very minor altercation . "

Saltsman would not speak about the possible use of an accelerant to start the fire or how the fire apparently spread so quickly.

" Even though we sent something to the lab does n't mean we will get a positive result either ," Saltsman said .

He said that a natural gas leak in the kitchen and tiki torches and a charcoal grill on the front porch had been ruled out as causes .

The tiki torches actually were golf-ball washers , he said .

Three residents of the house played on the Ohio State golf team .

As students walked by the house yesterday , many averted their eyes and tried to hold back tears.

Others , such as Angie Utz of Lebanon , Ohio , stopped to hug her friends .

Utz knew most of the victims and attended the party .

" These boys do n't start fights .

They did n't have any enemies ," Utz said .

Anger on the street began to brew after officials confirmed that the cause was arson and started passing out reward signs .

A disgusted Brian Connell , who lives two doors from the burned house , stapled up three signs .

" We knew it all along .

They were our friends , and we want to find out who did this ," Connell said .

Central Ohio Crime Stoppers , the Ohio Blue Ribbon Arson Committee , OSU and NorthSteppe Realty , the property manager , each contributed $5,000 to the reward fund .

Ohio University student leaders are interested in contributing , OU spokeswoman Leesa Brown said .

Dispatch reporter Alice Thomas contributed to this story

C.53 olsen background

Federal authorities took a suburban Spokane man into custody Wednesday on suspicion that he was trying to make a deadly biological weapon, and court papers indicate he that might have been targeting his wife.

The FBI arrested Kenneth R. Olsen, a 47-year-old high-tech worker from Spokane Valley, after agents determined that he had manufactured ricin, a poison made from castor beans, at his work cubicle at Agilent Technologies last summer.

On Wednesday, they searched his home, vehicle and his new workplace.

A criminal complaint filed Wednesday in U.S. District Court for the Eastern District of Washington accused Olsen of knowingly producing a biological agent, toxin or delivery system for use as a weapon.

Last summer, a fellow employee found bomb-making documents on a company printer that were eventually linked to Olsen, the complaint said. An investigation by Agilent officials determined that Olsen had searched the Internet for information on undetectable poisons.

According to the criminal complaint, Agilent officials fired Olsen on Aug. 30 and searched his cubicle, finding a printout titled"How to Kill" plus test tubes, a metal cup, a plastic bag with several brown seeds and a glass jar that contained a residue later determined to be ricin.

Olsen's cubicle also contained a paper on which Olsen had factored up a formula that would deliver a lethal dose of ricin to someone weighing 150 pounds, according to the complaint. Olsen's wife, Carol Olsen, weighs about 160 pounds, the complaint noted.

During an exit interview in August, Olsen"admitted to searching the Internet for information on explosives and poisons which he claimed pertained to Boy Scout projects he was researching," the complaint said.

Although authorities suspected in November that the substance was ricin, they didn't get final confirmation until recently, said FBI Special Agent Ray Lauer in Seattle.

Recipes for ricin abound on the Internet. The poison, said to be twice as deadly as cobra venom, can be made in days from castor beans, which are grown worldwide. It can be sprayed in the air or mixed with liquids or foods.

C.54 olsen new article

A Spokane Valley Scoutmaster and father of four who was arrestedafter federal authorities accused him of making one of the most toxic natural poisons known was probably planning to use it on his wife , authorities said yesterday .

Kenneth R. Olsen , 47 , was arrested in Spokane yesterday morning **.**

He may have intended to poison his wife after he had an affair with another woman , federal court papers say .

He was arrested on a federal complaint alleging that he knowingly produced and possessed a biological agent for use as a weapon .

Authorities provided no further details .

The poison is ricin , which is easily derived from castor beans **.**

While traces of ricin have been found at suspected al-Qaida sites in Afghanistan and in some domestic terrorist and murder schemes over the last decade , there is no sign that Olsen case is connected with terrorists , federal agents said .

" There was not a terrorist connection per se , no mass murders planned ," said FBI spokesman Ray Lauer of the Seattle office .

In Spokane , FBI agent Norm Brown said : " We have no knowledge he had any connections with outside or national organizations .

In our opinion , the public has nothing to fear from this incident . "

Olsen said little during a preliminary appearance in federal court in Spokane.

A detention hearing was set for Monday , and Olsen will remain in jail until then .

Olsen attorney , John Clark , said his client is not guilty of any crime.

" He is a Scoutmaster , active in his church and a computer and Internet hobbyist ," Clark said .

" I assume that 9/11 has caused the federal government to look at everything as if it was a real threat ."

According to court documents , Olsen said he was " searching the Internet for information on explosives and poisons which he claimed pertained to Boy Scout projects he was researching . "

Olsen came under investigation initially in August by **the Spokane County Sheriff Department**, which turned the case over to the FBI **.**

Brown said Olsen arrest was delayed because the FBI has been swamped with work by the Sept. 11 terrorist attacks .

The investigation began when Olsen co-workers grew worried after finding items at his workstation , documents say .

Papers found in his cubicle at the Agilent Technologies Inc. plant **in Liberty Lake suggested**that Olsen was trying to determine how much ricin was needed to kill a 150-pound person , the approximate weight of his wife , court documents said **.**

" Ricin is the second-most deadly toxin known to man ," assistant U.S. attorney Earl Hicks said in court .

" This is a very serious case . "

Neighbors near Olsen suburban Spokane home were not considered in danger , Brown said .

Olsen wife and three children were removed from the home while agents searched it .

A fourth child no longer lives at home .

Neighbor Dave Rausch said the family appeared normal , although they are private.

" Certainly it scary ," he said .

" You would n't think something like this would take place in a neighborhood like this. "

Congress in November toughened sentencing guidelines for chemical and biological weapons.

Though spurred by the Sept. 11 attacks , the stiffer guidelines were the result of years of efforts triggered by a 1995 sarin nerve gas attack on a Tokyo subway that killed 12 people , and after several cases of domestic terrorism and homicides involving ricin in the 1990s .

Ricin is a natural poison that has both criminal and , apparently , medicinal applications.

On the one hand , ricin is an untraceable biological weapon that is twice as deadly as cobra venom **, though not as lethal as botulism .**

On the other hand , ricin has been referred to as a potential " magic bullet " against cancer.

Castor beans are grown all over the world and the toxin is relatively easy to produce .

It can be used to poison water or food , sprayed into the air or injected into a person **; it can kill within three days of exposure .**

In very small doses , it causes the human digestive tract to convulse - hence the laxative effect of castor oil .

But in larger doses , ricin causes diarrhea so severe that victims can die of shock , as a result of massive fluid and electrolyte loss .

According to 1999 testimony by the FBI before Congress , authorities have documented several domestic criminal cases involving ricin , including these **:**

In 1995 , four members of the Minnesota Patriots Council , an extremist tax-protest group with anti-government ideals , were arrested in connection with a plot to kill a U.S. marshal with ricin .

In 1993 , Thomas Lavy , an Alaskan white supremacist entering Canada on his way to North Carolina , was arrested after trying to cross the border with what authorities said was $100,000 cash , a gun and a small container of ricin .

The FBI later found a large quantity of ricin in his home .

Lavy committed suicide while in detention .

This report contains information from The Associated Press

C.55 outreach background

A judge has ordered a couple and their son's babysitter to stand trial for allegedly killing the 9-year-old boy, who authorities say suffocated after being wrapped head to toe with duct tape for stealing food.

After a daylong hearing Thursday, Judge John Bennett found sufficient evidence to try the three for the Dec. 30 death of Brian Edgar.

Neil Edgar Sr., 47; Christy Edgar, 46; and Chasity Boyd, 19, also face charges for abusing two of their other children. The Edgars also have two sons, ages 16 and 12, and a 9-year-old daughter.

The Edgars are pastors of God's Christian Outreach Ministry in Kansas City, Kan., and Boyd is a member of the storefront church. Five other church members face charges of abusing Brian Edgar.

Attorneys for the Edgars and Boyd entered innocent pleas for them on all counts.

A gag order has been issued in the case. Robert Kuchar, an attorney for Boyd, said that while he could not discuss his client's case in detail, "We look forward to the opportunity to present a full picture at a jury trial."

Prosecutors alleged that the Edgars and Boyd systematically abused the Edgars' children, who were all adopted, as punishment for taking food or sneaking water from the kitchen faucet.

The 16-year-old son testified that his parents were punishing Brian for stealing food by wrapping him"like a mummy" with duct tape and had even gone to the store to buy six more rolls of tape.

The teen told the court that on the night of Dec. 29, his mother and Boyd had wrapped Brian in duct tape until only his nose showed, and then Boyd carried him into a small storage room where he was placed on a sleeping bag.

Brian wasn't breathing when his parents checked on him the next morning, and his father took him to the University of Kansas Medical Center, where he was pronounced dead.

Dr. Richard Dietz, an emergency room physician, testified that Neil Edgar said he had given Brian an herbal sleep aid containing melatonin to keep him from waking up and stealing food, and asked if melatonin could have killed his son. The doctor said he answered no; an autopsy found no trace of melatonin.

C.56 outreach new article

A judge on Thursday ruled that there is enough evidence to send a Kansas City , Kan. , couple and their baby sitter to trial on charges they killed their 9-year-old son .

Neil Edgar , 47 , his wife , Christy Edgar , 46 , and their baby sitter , 19-year-old Chasity Boyd , are charged with first-degree felony murder in the death of Brian Edgar .

Their preliminary hearing was Thursday in **Johnson County District Court**.

The Edgars , who had adopted Brian , were pastors with God Creation Outreach Ministry in Kansas City , Kan.

They had rented a house in Overland Park , where authorities allege Brian was killed .

Friends and relatives of the Edgars and church members filled the courtroom Thursday morning .

After hearing testimony , Judge John Bennett determined there was enough evidence for the Edgars and Boyd to stand trial **.**

A trial date was expected to be set Friday .

The hearing included testimony from physician Richard Deitz , who described the morning of Dec. 30 when Neil Edgar Sr. carried Brian into the emergency room at University of Kansas Medical Center **.**

As emergency room personnel tried to revive the boy , it " became very apparent that young Brian was dead ," Deitz testified Thursday .

The onset of rigor mortis indicated that Brian had been dead for quite a while , Deitz testified .

Brian died from asphyxiation after being bound by duct tape , prosecutors said **.**

Deitz said he walked to the waiting room where Neil Edgar was sitting and told him the boy was dead .

Edgar put his head in his hands , began to cry and repeatedly said , " Oh Jesus , oh God , what have I done ," Deitz testified .

Deitz asked Edgar what he meant .

Edgar told him that Brian had been waking up in the night and " stealing food ," Deitz testified .

Edgar said he had given Brian one dose of an herbal sleep aid that contained melatonin , and asked the doctor if he thought that could have caused Brian to die , Deitz testified **.**

Deitz said he told Edgar that he did n't think one pill would do that.

Erik Mitchell , the pathologist who performed an autopsy on Brian , on Thursday described the adhesive marks around Brian head .

He said some of the marks on Brian wrist and ankles appeared to be old and some very recent .

Mitchell testified that whatever caused those marks had been repeatedly used over time

---

C.57 rblake background

Blake Pleads Not Guilty to Murder;

Before actor Robert Blake pleaded not guilty Thursday to killing his wife, he sp oke to reporters for the first time since he was released from jail two weeks ag o, thanking the judge who set his bail for"saving my life" and calling himself t he"luckiest person you will ever meet." :

Inside the courtroom at his formal arraignment, Blake, who looked stronger and s aid he had gained 12 pounds, entered a not guilty plea to murder, two counts of soliciting murder and conspiracy, and waived his right to a speedy trial to give his lawyer more time to investigate the case.

Attorney Thomas A. Mesereau Jr. said he plans to ask the judge to appoint a spec ial master to oversee forensic testing of police evidence by a defense expert.

Los Angeles County Superior Court Judge Darlene Schempp ordered Blake and co-def endant Earle S. Caldwell back to her Van Nuys courtroom on June 19 for a pretria l hearing on motions, including a routine defense motion to dismiss the case.

The former"Baretta" star is accused of fatally shooting Bonny Lee Bakley, 44, ne arly two years ago as she sat in his car near the Studio City restaurant where t hey had dined. He also is accused of soliciting two stuntmen and conspiring with Caldwell to kill her. He faces life in prison if convicted.

Outside the courthouse, the 69-year-old actor told reporters that he has been ea ting and sleeping a lot since he walked out of Men's Central Jail on March 14, a fter being held without bail for 11 months. A condition of his release on $1.5- million bond was that he be placed under house arrest and hooked up to an electr onic monitor device to ensure that he remains at home.

"I believe in the system," he said during the impromptu news conference. "I am s tanding here in front of you and that is proof that this is still America and it still works."

Blake did not answer a question about whether he had seen his youngest daughter, Rose, but said his family was"safe and healthy and happy."

He thanked Los Angeles County Sheriff Lee Baca and his staff"for doing their bes t to keep this old, busted-up cowboy on the right side of the dirt." And he cred ited Judge Lloyd M. Nash, who presided over his preliminary hearing, with"saving my life," apparently for setting bail.

"I am the luckiest person you will ever meet if you live to be a million," the a ctor said. "I've had everything there is to have in life. If I was going to chec k out, it wasn't going to be with any great regrets, but I'm glad I didn't." GRA PHIC: PHOTO: OUT ON BAIL: Robert Blake is escorted into courthouse by his attorn ey, Thomas A. Mesereau Jr., right, and bodyguards. "I believe in the system," he told reporters. "It still works." PHOTOGRAPHER: Anacleto Rapping Los Angeles Ti mes PHOTO: NEXT SCENE: Robert Blake, accused of shooting Bonny Lee Bakley to dea th two years ago, sits in Van Nuys courtroom for his arraignment Thursday. He ap peared stronger and said he had gained 12 pounds since his release on bail two w eeks ago. PHOTOGRAPHER: Pool Photo

C.58 rblake new article

Robert Blake trial expected in October ; actor pleads innocent

Appearing in court for the first time since he was allowed to post bail March 14,a healthier looking Robert Blake pleaded innocent to murdering his wife and waived his right to a speedy trial until October , when proceedings are likely to begin .

The star of the old " Baretta " TV detective series told reporters he has gained about 12 pounds and has been sleeping a great deal .

" I want to thank Judge (Lloyd) Nash for saving my life ," Blake said of the judge who granted him release on $1.5 million bail and ordered him to trial .

He is confined to his residence with electronic monitoring .

Blake will be tried on charges of murdering Bonny Lee Bakley , solicitation of murder , conspiracy and the special circumstance of lying in wait .

Superior Court Judge Darlene Schempp set a June 19 date for pretrial motions .

Also at Thursday arraignment hearing , Blake former handyman-bodyguard , Earle Caldwell , pleaded innocent to a murder conspiracy charge .

Prosecutors filed documents compiled from testimony during the defendants ' preliminary hearing , which showed that the case against Blake depends almost entirely on the testimony of two aging stuntmen who said Blake solicited them to murder his wife .

Both men said they refused and suggest Blake then took the matter into his own hands .

The 69-year-old actor is accused of killing Bakley on May 4, 2001 , after the pair dined at Vitello restaurant , Blake longtime hangout in his Studio City neighborhood .

Bakley , 44 , who became Blake wife after giving birth to his baby girl , Rosie , was shot in the head and upper body as she sat in Blake car outside the restaurant .

Blake claims he found his wife mortally wounded after he went back to the restaurant to retrieve a handgun he carried for protection .

Prosecutors say Blake despised Bakley , a con artist with a criminal record , and wanted to find a way to get rid of her but keep their baby .

Blake has suggested she was killed by a victim of one of her con schemes .

A key piece of evidence for the prosecution is a prepaid telephone card on which Blake allegedly made 56 calls to one of the stuntmen allegedly solicited as a killer .

Prosecutors also say Blake withdrew $126,000 from one of his bank accounts in the months before the killing .

The new allegations list 38 overt acts , only four of which mention Caldwell , who is free on $1 million bail posted by Blake last year .

Prosecutors have characterized a list found in Caldwell car as a shopping list for murder .

It includes such items as shovels , a sledge , duct tape , lye , pool acid and the notation : " Get blank gun ready . "

His lawyer has said the list includes typical handyman items .

GRAPHIC : AP Photos CANU107 ,

C.59 robichaud background

A former Roman Catholic pastor accused of raping a teenage boy in 1985 apologizes to his accuser in a secretly recorded conversation, saying he was drinking at the time and"played affectionately."

"I hated you, I hated you so much," the accuser, now a 33-year-old state trooper, is heard saying in the tape played for jurors Wednesday.

"I know you never meant to hurt me," said the trooper, whom police had wired with a recording device for his April 2002 meeting with the priest.

The tape was played during the trooper's second day of testimony at the Belknap County Superior Court trial of the Rev. George Robichaud, 58. Robichaud is the first priest to face criminal sex assault charges in New Hampshire since the church-abuse scandal erupted 18 months ago.

Robichaud has admitted to inappropriate sexual contact with the former altar boy, who regarded him as a father figure. Robichaud has pleaded innocent to rape and attempted rape. He has been on leave since he was accused last year.

His lawyer denies Robichaud admitted guilt in the recorded conversation and questions the trooper's recollection of his age at the time. The trooper has said he was 14, 15 or 16. The age of consent in New Hampshire is 16.

"I'm going by what I believe happened in my life. I'm here to tell the truth," the trooper told defense lawyer Peter Callaghan.

"But you still have doubts," Callaghan said.

In his taped meeting at the Wolfeboro church where Robichaud was the pastor, the trooper discussed the clergy sex-abuse scandal then receiving wide news coverage. He said, "some memories have come up that have been really tough to deal with" regarding his relationship with Robichaud.

"I'm not looking for any type of revenge or retribution," the trooper said. "I am looking for an apology from you for the things that happened between us."

Robichaud responded, "Let me say something. You know I was drinking. I didn't mean to do anything to hurt you. ... I played affectionately, um, I am sorry."

Robichaud said, "Now let me say this. I went to a therapeutic program, dealt with a lot of issues, OK? (inaudible) ... and even better understand my sexuality. And I'm sorry."

After the tape was played, the trooper testified that he didn't hate Robichaud. "I still have a lot of compassion for him because there were a lot of good times in my life with him."

Robichaud faces 7 1/2 to 15 years in prison on the rape charge under sentencing provisions in effect in 1985. The same crime today carries a sentence of 10 to 20 years.

The man said Robichaud befriended him in 1984 when the priest was pastor of of St. Anthony's and St. Stephen's churches in Swanzey. He visited Robichaud in Rome, where the priest was attending a conference at the Vatican.

He stayed for seven days in an adjoining room at a convent and said that Robichaud assaulted him, lying on top of him fully clothed.

In the spring of 1985, Robichaud took the boy to his cottage on Lake Winnisquam, fondled him and had sex with him briefly, the trooper testified. The charges resulted from that incident.

C.60 robichaud new article

LACONIA –

A former altar boy who is now a New Hampshire police officer testified in Belknap County Superior Court yesterday about how he was befriended by his Catholic priest , taken on trips , groped , then raped in **the summer of 1985**.

But a defense attorney for the Rev. George H. Robichaud , 58 , of Sanbornton , told the jury the case hangs on whether the 33-year-old alleged victim was younger than 16 when the incident occurred .

If he was 16 or older , the age of consent , it would not be considered a crime under state law .

And he also noted his client did not admit to sexual penetration in a secretly taped meeting the two had last spring , which was approved by **former Attorney General Phillip T. McLaughlin** .

Concord attorney Peter G. Callaghan said there is no physical evidence in the case , only the words of a man who is not sure whether the events occurred prior to his 16th birthday .

The officer secretly taped a conversation last spring , in which Robichaud allegedly apologizes for inappropriate sexual contact .

The jury is expected to hear the tape today when the trial resumes **.**

The victim said he feels confident that Robichaud sexually penetrated him while he pretended to sleep in 1985, when he was alone as a guest at Robichaud cottage on the shores of Lake Winnisquam **.**

He said he believes that because his life became busier after that time with sports , school , jobs and friends , that it was unlikely he would have been able to have gone with " Father George " on an overnight trip .

As the scandal involving the Catholic Church boiled over last year , the buried thoughts of the assault resurfaced **, he said , and he felt it was his duty as a police officer , father and citizen to come forward with the allegations .**

It is this newspaper policy not to reveal the names of victims of sexual assault **.**

Recently , another police officer , State Police Sgt. Phil Jepson announced that he was molested by a teacher at Bishop Guertin High School in the 1980s .

The teacher denied the accusations .

Robichaud **, of 284 Black Brook Road ,**Sanbornton , has pleaded innocent to one count of felonious sexual assault and one count of attempted aggravated felonious sexual assault in the case **.**

A second and separate set of indictments were handed down in January, alleging raped another altar boy from Laconia between June 1 and Oct. 28, 1982 , at the cottage.

Robichaud has denied those charges , as well as those in a civil case alleging between 20 and 30 assaults .

Robichaud has been a priest at St. Cecilia Roman Catholic Church in Wolfeboro and St. Joan of Arc parish in Alton , but the charges that are now going forward stem from his time at St. Anthony and St. Stephen churches in Swanzey during the 1980s .

Robichaud is now on administrative leave **.**

Belknap County Attorney Lauren Noether , in opening statements , said Robichaud " broke a sacred trust " and that he knowingly used his authority to coerce the boy to submit .

Callaghan said his client enjoys the presumption of innocence and the state has the burden of overcoming that presumption beyond a reasonable doubt .

In the first day of **what is expected to be a five-day trial, the primary witness took the stand but was not cross-examined .**

Born in the Keene area and adopted by a local family as an infant , he said his mother was a devout Catholic and his father was an alcoholic who rarely went to church .

His mother encouraged his relationship with Robichaud , which began as visits to the rectory after Mass when he was 13 and 14 , and included lobster dinners and occasional trips to the cottage on Lake Winnisquam .

Not only was Robichaud a church leader and role model , he became a sort of surrogate father who showered affection on him , something he admits he was starved for as a boy .

He recalled only once getting a hug from his father in those years **.**

" I 'd get hugs from him (Robichaud) where I was n't getting them from my dad . It felt nice ," he said , wiping away a tear .

There was a noticeable change in the relationship when he was 14 and was asked by Robichaud to lie on his bed at the rectory **.**

It was there , while a nun was in the room next door watching television, that he said Robichaud lay on top of him , fully clothed **, and kissed him and moved his hands and hips about the boy body .**

Asked by Noether why he did n't tell anyone he said , " There was nobody really to tell ," he said .

His mother , who worked on the books at the church , encouraged the relationship.

The alleged victim said his family was fairly poor and Robichaud offered him experiences he could not get from his family.

He said the priest let him drive his Jeep before he had a license , allowed him to drink at the cottage and he learned to drive his power boat .

In the fall of 1984 **, Robichaud even offered him an all-expenses-paid trip to Italy and**he said that during the trip , when they had adjoining rooms at a convent outside the Vatican , Robichaud entered his bedroom and groped him in his boxer shorts and kissed him **.**

But the two never spoke of the incident and he did not sever his relationship with either Robichaud or the church for some time because he felt welcome and appreciated .

During the summer of 1985 , he believes , Robichaud sexually penetrated him while he was dozing **on the couch at the house in Sanbornton at night.**

He described the priest as being " slow , sneaky , trying not to wake me up ," but that he knows what he experienced , though the two never spoke of that incident until last year

---

C.61 valentine background

High school theology teacher fired after giving student'I wish you

A theology teacher at a Roman Catholic high school lost his job after giving a student a valentine that read"I hate you, I wish you would die."

A police report stated R. Scott Jones passed out similar cards Feb. 14 to his other students at St. Mary's High School, many of whom regarded it as a joke, but one 17-year-old boy said he was"freaked out" by the card.

According to the report, Jones handed the student the card and said"I made this for you."

During class, the report stated, the teacher used an eraser to write the word"Die" on the chalkboard while looking at the boy, then smiled.

Police said the incident remains under investigation. Jones, 44, was placed on administrative leave. He wouldn't comment.

"A teacher is in a position of authority," said Kim Sue Lia Perkes, a spokeswoman for the Roman Catholic Diocese of Phoenix. "It's not right to do something like that to a student. You can't even make a joke about something like that."

C.62 valentine new article

High school teacher fired after penning " Die , Die , Die " valentine

A teacher at a Catholic high school here lost his job after giving a student a valentine that **included the words , " Die , Die , Die .**"

As of Monday , 44-year-old R. Scott Jones was no longer employed at St. Mary High School .

The theology teacher had been placed on administrative leave following the Feb. 14 incident .

" A teacher is in a position of authority ," said Kim Sue Lia Perkes , a spokeswoman for the Roman Catholic Diocese of Phoenix .

" It not right to do something like that to a student .

You ca n't even make a joke about something like that . "

Jones declined comment , although the police report noted that Jones passed out similar cards to other students in his class , many of whom regarded it as a joke .

But Jones ' message startled one 17-year-old boy who he said he was " freaked out " after receiving the card .

According to the report , Jones handed the student a card on Valentine Day and said , " I made this for you . "

When the student opened the card he found the message : " I hate you , I wish you would die , Happy V Day , Die , Die , Die . "

During the class , the teacher allegedly wrote on the chalkboard with the eraser the word , " Die ," while looking at the boy , then smiled .

After class , the student went to the front office and told school officials what happened .

Police were called .

Officers tried to talk to Jones that day , but they were told he left early .

When they went to his classroom , an officer noted that the chalkboard still had the " Die " message that the student had described .

Police say the matter remains under investigation

Appendix D

Additional Lexicon Experiments

This chapter contains the remaining tables from Chapter 8 from the complete range of experiments on the different options, such as the choice or lexicon or restraints on the words to use in forming co-occurrences.

Lexicon	Precision	Recall	F-measure
Dekan	0.690	0.766	0.711167922497309
Empty	0.695	0.789	0.720761041009464
Minimal	0.689	0.815	0.722510293360782
Combine	0.686	0.772	0.709718574108818
Nominals	0.697	0.800	0.725003250552594
Morph	0.687	0.824	0.723065525609912

Table D.1: All words are used, with both the promiscuity and document frequency options on, but the docsim feature is not used. Precision does not vary much across the different lexicons, but recall does. The result is interesting because the Dekan and Empty lexicons do relatively well here.

Lexicon	Precision	Recall	F-measure
Dekan	0.683	0.759	0.704152404237979
Empty	0.686	0.813	0.719728997289973
Minimal	0.694	0.796	0.721745492552914
Combine	0.692	0.775	0.714971337155046
Nominals	0.688	0.804	0.719126365054602
Morph	0.693	0.818	0.726295964125561

Table D.2: Adjectives are removed from consideration, but all other content words are used here. Both the promiscuity and document frequency options are on, and the docsim feature is not used. The results are more typical, with the more sophisticated lexicons performing better.

Lexicon	Precision	Recall	F-measure
Dekan	0.681	0.775	0.706715318693091
Empty	0.695	0.786	0.720007908264136
Minimal	0.685	0.816	0.719660100424874
Combine	0.696	0.789	0.721513598738668
Nominals	0.695	0.800	0.7234873129473
Morph	0.688	0.791	0.715968951453756

Table D.3: Only head words are used, and neither the promiscuity or document options are on. The docsim is not used. In this experiment and the next, also using head words only, the Empty and the Combine lexicons do relatively well, likely because with fewer words, the system is more senstive to noise.

Lexicon	Precision	Recall	F-measure
Dekan	0.676	0.784	0.705141032464077
Empty	0.697	0.770	0.717404090362251
Minimal	0.698	0.777	0.719960175228992
Combine	0.693	0.774	0.71546218487395
Nominals	0.687	0.772	0.710467515070328
Morph	0.699	0.768	0.718362103572862

Table D.4: Only heads are used, and both the promiscuity and document options are used; the docsim feature is not used.

Lexicon	Precision	Recall	F-measure
Dekan	0.698	0.755	0.714175362515246
Empty	0.691	0.786	0.71699801980198
Minimal	0.696	0.822	0.729548584544759
Combine	0.701	0.797	0.727280656079146
Nominals	0.684	0.805	0.716300247170548
Morph	0.693	0.808	0.723909502262443

Table D.5: All words, with promiscuity but no document frequencies

Lexicon	Precision	Recall	F-measure
Dekan	0.693	0.765	0.713135593220339
Empty	0.686	0.783	0.712479108635098
Minimal	0.708	0.809	0.735548991909593
Combine	0.683	0.776	0.708472129394466
Nominals	0.684	0.805	0.716300247170548
Morph	0.683	0.819	0.718808789514264

Table D.6: All words, with document frequencies but no promiscuity

Lexicon	Precision	Recall	F-measure
Dekan	0.704	0.806	0.731782305906629
Empty	0.724	0.805	0.74653516075317
Minimal	0.723	0.845	0.755733547748639
Combine	0.726	0.823	0.75261116009573
Nominals	0.725	0.836	0.755076616419584
Morph	0.726	0.841	0.757056416615003

Table D.7: All words, both promiscuity and document frequency options on, and with the docsim feature used.

Lexicon	Precision	Recall	F-measure
Dekan	0.713	0.799	0.736791257113295
Empty	0.719	0.822	0.747083807356845
Minimal	0.721	0.827	0.749832746478873
Combine	0.721	0.821	0.748345132743363
Nominals	0.727	0.843	0.758303637713437
Morph	0.731	0.854	0.764011748867948

Table D.8: Heads only, with the docsim feature, without promiscuity or document frequency.

Lexicon	Precision	Recall	F-measure
Dekan	0.711	0.789	0.732731191222571
Empty	0.723	0.823	0.750351828499369
Minimal	0.719	0.796	0.740489067149696
Combine	0.716	0.829	0.746527480820023
Nominals	0.722	0.825	0.750094446543257
Morph	0.731	0.826	0.757123510971787

Table D.9: Heads only, with the docsim feature, with both promiscuity and document frequency options

Appendix E

Newsblaster Cluster

E.1 BBC March 23

A bomb tore through a shopping centre in Lebanon's anti-Syrian Christian heartland north of Beirut killing three people, including two foreign workers.

It was the second blast in a Christian area in days, sharpening fears of sectarian chaos weeks before elections.

The explosion at the Kaslik shopping centre occurred at about 0130 local time (2330 GMT), police said.

Shop windows were shattered and glass littered streets lined with boutiques, jewellery stores and nightclubs.

Lebanon has been plunged into political turmoil since the assassination of ex-Prime Minister Rafik Hariri on 14 February.

President Emile Lahoud has ordered an investigation into the attack in Kaslik, near the port of Jounieh, about 15km (10 miles) north of Beirut. He said it aimed to drive Lebanon into "chaos and fear" and he renewed calls for talks between opposition and loyalist politicians "as the only means to break the current deadlock and bridge all differences".

Political message

The bomb is thought to have been left in a leather bag at the back entrance of the shopping centre, a Lebanese security official said.

Two of the three fatalities were Indians, with the nationality of the third still to be determined, police said.

The roof of the centre - which was closed at the time of the blast - collapsed, and local television showed rescuers searching through the rubble.

One body was shown retrieved and covered with a blanket.

One Lebanese person was also reported to have been injured in the blast.

"It is clear that those who carried out this attack are targeting the security and stability of the country," opposition lawmaker Faris Bouez told reporters at the scene.

"It is a political message to the [anti-Syrian] independence uprising," he said.

"We don't know what is happening, but it's obvious that we are very frightened," said local resident Claude Boustani.

The opposition has blamed Damascus supporters for recent violence, saying they are keen to stir unrest to justify the presence of Syrian troops in Lebanon.

At about midnight on Friday, another blast took place in the northern suburb of New Jdeideh, a part-residential, part-commercial area, injuring 11 people.

Following the pressure on Damascus after Mr Hariri's assassination, some Syrian troops are now withdrawing from the country.

Demonstrations and counter-demonstrations, although largely peaceful, have kept tension high between Lebanon's pro-and anti-Syrian camps.

E.2 ABC March 24

A U.N. report into the assassination of former Lebanese Prime Minister Rafik Hariri concluded that Lebanon's probe of the killing was riddled with flaws and an international investigation is needed.

The report, released Thursday, does not directly assign blame, saying the causes could not be determined. But it says Syrian military intelligence shares responsibility to the extent that it and Lebanese security services failed to provide "security, protection, law and order" in Lebanon.

The report says there was a "distinct lack of commitment" by Lebanese authorities to investigate the crime, and the probe was not carried out "in accordance with acceptable international standards."

It detailed a host of problems, including the disappearance of crucial evidence and tampering with the scene of the massive bombing that killed Hariri. The report even faults police for not turning off a water main that flooded the blast crater and washed away vital evidence.

In Beirut, Lebanese President Emile Lahoud responded by saying he had told U.N. Secretary-General Kofi Annan to do "what is necessary" to learn who was behind the Feb. 14 killing.

Hariri died in a blast in central Beirut that killed 17 other people. The Lebanese opposition has blamed Syria and its Lebanese allies, who have both denied any involvement.

"It is clear that the assassination took place in a political and security context marked by an accute polarization around the Syrian influence in Lebanon," the report said.

The opposition and Hariri's family have insisted on an international investigation, saying they have no trust in the Lebanese probe. The report reflected that sentiment, saying the Lebanese investigation "lacks the confidence of the population necessary for its results to be accepted."

In Washington, the State Department supported the recommendation for an international commission to investigate the attack.

"The report once again makes clear the importance of immediate and full withdrawal of all Syrian military and intelligence forces from Lebanon, in accordance with U.N. Security Council Resolution 1559," Deputy State Department spokesman Adam Ereli said. "The Lebanese people deserve a government capable of leading them forward to prompt, free and fair elections, without foreign interference and in the presence of international observers."

E.3 Boston Globe March 24

Lebanon's inquiry into the killing of former Prime Minister Rafik al-Hariri was seriously flawed and an independent investigation is needed to "find the truth," a U.N. fact-finding team said on Thursday.

In what could be the most damning piece of evidence, the team's report gave credence to alleged threats made at a meeting of "physical harm" by Syrian President Bashar al-Assad to Hariri prior to his Sept. 8 resignation as Lebanon's prime minister.

The report cited numerous accounts of the meeting between the two based on Hariri's statements to others. They had met to discuss extending the term of Lebanon's Syrian-backed President Emile Lahoud, which Hariri and Druze opposition leader Walid Jumblatt opposed.

Assad was quoted as saying he "would rather break Lebanon over the heads of Hariri and Jumblatt than see his word in Lebanon broken."

In response, however, Syria's U.N. Ambassador Fayssal Mekdad said he was quite certain Assad "did not threaten physical harm."

The fact-finding mission, led by Irish Deputy Police Commissioner Peter Fitzgerald, said Syrian military intelligence bore primary responsibility for a lack of security, protection and law and order and that and Lebanese security forces showed "systematic negligence."

"It became clear to the mission that the Lebanese investigation process suffers from serious flaws and has neither the capacity nor the commitment to reach a satisfactory and credible conclusion," Fitzgerald wrote.

The United States and France were expected to introduce a resolution in the U.N. Security Council calling for an international inquiry, council diplomats said.

The State Department said the fact-finding report raised "serious and troubling allegations" and said the United States wants an independent, international commission to conduct an investigation into the matter.

"The report once again makes clear the importance of immediate and full withdrawal of all Syrian military and intelligence forces from Lebanon," State Department deputy spokesman Adam Ereli said in a statement. SYRIA BLAMES U.S., FRANCE FOR DIVISIONS

"I expect the council to support the idea that there should be an independent investigation," British Ambassador Emyr Jones Parry told reporters.

Lahoud urged the United Nations to "do what's necessary to reveal the truth in the crime," the Lebanese presidential palace said in a statement.

Syria, for its part, reiterated denials that it played any role in the killing and said that Lebanese authorities were able to investigate it on their own.

But Mekhad said, "It is up to them, we don't interfere with them in their affairs."

Syria had created "a peaceful atmosphere" in Lebanon while the United States and France had caused divisions there by calling for Syria to withdraw forces in a Security Council resolution adopted last Sept. 2, Mekdad told reporters.

U.N. Secretary-General Kofi Annan has said he expects Syria to complete its withdrawal before May elections in Lebanon, but while Damascus has agreed to pull out, it has yet to publicly set a timetable for its departure.

The U.N. Security Council ordered the fact-finding mission to the region last month to report on "the circumstances, causes and consequences of the assassination."

Hariri was killed in a Feb. 14 bomb attack on his motorcade in Beirut. Lebanon's anti-Syrian opposition and many ordinary Lebanese have pointed a finger at Syria and its local allies.

Opposition leaders had demanded an international investigation into the killing, saying they did not trust pro-Syrian Lebanese security chiefs.

The U.N. team said it thought the blast that killed Hariri may have been caused by a suicide bomber in a 1995 or 1996 Mitsubishi truck, carrying a TNT charge of about 2,200 pounds (1,000 kg).

Before the killing, despite widespread rumors that Hariri and Jumblatt were in danger, "none of the security services had taken additional measures to protect any of them," the U.N. report said.

After the killing, the security services removed some evidence from the scene and falsified or destroyed other evidence rather than secure the area, it said.

The U.N. mission said the consequences of the killing could be far-reaching and that "Lebanon could be caught in a possible showdown between Syria and the international community, with devastating consequences for Lebanese peace and security."

E.4 Boston Globe March 24

UN Secretary General Kofi Annan said yesterday a more in-depth investigation may be needed into the killing of former Prime Minister Rafik al-Hariri of Lebanon than a UN report due to be released shortly.

In a speech to Arab leaders, including President Bashar Assad of Syria, Annan called for free and fair parliamentary elections in Lebanon in May.

"Within the next few days, I expect to release the report of the mission of inquiry I established in the wake of the killing. A more comprehensive investigation may well also be necessary," Annan told an Arab League summit.

He told reporters before leaving Algiers that the team, headed by Irish police commissioner Peter Fitzgerald, would report to him today.

But he added: "Often in these cases you are going to want a broader investigation, more than the fact-finding team."

Diplomats in New York expect Fitzgerald to recommend an international investigation, and said the United States and others wanted to ask the 15-member Security Council to authorize a probe into the Hariri killing.

Annan spoke to the summit shortly after a bomb tore through a shopping center in an anti-Syrian Christian heartland north of Beirut, killing three people and bringing Lebanon closer to chaos ahead of elections.

Hariri was killed Feb. 14 in a bomb attack on his motorcade in Beirut.

Lebanon's opposition and many ordinary Lebanese have pointed a finger at Syria and its local allies. Members of the anti-Syrian opposition have demanded an international investigation into the killing, saying they do not trust pro-Syrian Lebanese security chiefs. Syrian and Lebanese officials have denied any role.

The Lebanese judge in charge of a local investigation into Hariri's assassination, Michel Abu Arraj, asked yesterday to be taken off the case, judicial sources in Lebanon said. They said Lebanon's judiciary would decide after a meeting today whether to accept the request and appoint another judge to investigate.

Annan said on Tuesday after talks with Assad in Algiers that he expected Syria to present a credible and precise timetable on a full withdrawal of its troops and security services from Lebanon by early next month.

E.5 Boston Globe March 24

A U.N. report into the assassination of former Lebanese Prime Minister Rafik Hariri concluded that Lebanon's probe of the killing was riddled with flaws and an international investigation is needed.

The report, released Thursday, says there was a "distinct lack of commitment" by Lebanese authorities to investigate the crime, and the probe was not carried out "in accordance with acceptable international standards."

It detailed a host of problems, including the disappearance of crucial evidence and tampering with the scene of the massive bombing that killed Hariri. The report even faults police for not turning off a water main that flooded the blast crater and washed away vital evidence.

In Beirut, Lebanese President Emile Lahoud responded by saying he had told U.N.

Secretary-General Kofi Annan to do "what is necessary" to learn who was behind the Feb. 14 killing.

Hariri died in a blast in central Beirut that killed 17 other people. The Lebanese opposition has blamed Syria and its Lebanese allies, who have both denied any involvement.

The report does not directly assign blame, saying the causes could not be determined. But it says Syrian military intelligence shares responsibility to the extent that it and Lebanese security services failed to provide "security, protection, law and order" in Lebanon.

"It is clear that the assassination took place in a political and security context marked by an accute polarization around the Syrian influence in Lebanon," the report said.

The opposition and Hariri's family have insisted on an international investigation, saying they have no trust in the Lebanese probe. The report reflected that sentiment, saying the Lebanese investigation "lacks the confidence of the population necessary for its results to be accepted."

Hariri's killing led to political turmoil in Lebanon. Mass demonstrations forced the resignation of the Lebanese government and intensified the international campaign for Syria to withdraw its troops from the country.

Syria has now pulled back its troops and intelligence agents into eastern Lebanon toward the border and has been promising to work out their complete removal with the pro-Syrian government in Beirut.

The report is from an investigation carried out by team led by deputy Irish Police Commissioner Peter Fitzgerald, appointed at the behest of the U.N. Security Council.

In a letter accompanying the report, Annan endorsed the recommendation for a new investigation.

In his report, Fitzgerald also faults Syria for interfering in the governing of Lebanon "in a heavy-handed and inflexible manner." He said his investigators also received testimony that Syrian President Bashar Assad had threatened Hariri and leading opposition figure Walid Jumblatt with physical harm.

Syria's U.N. Ambassador Fayssal Mekdad rejected the report, saying it contained "too much rhetoric." He again denied his country had any role in Hariri's assassination.

Fayssal called Hariri a "great ally of Syria" and instead blamed the U.N. Security Council for passing resolution 1559, which demands Syria's withdrawal from Lebanon.

"We think that things were going on well in Lebanon until a certain development that has taken place here in this building when one, two countries pushed the council to adopt a resolution that was not called for," Mekdad said from the United Nations in New York.

The report went on to demand an international independent commission with the authority to interrogate witnesses, conduct searches and other tasks. Fitzgerald said such an inquiry would be impossible without Lebanon's cooperation.

The U.N. Security Council may now take up the issue. It would have to approve a resolution seeking a new investigation, and would most likely ask Annan to appoint a new

team.

"I expect the council to support the idea that there should be an independent investigation," Britain's U.N. Ambassador Emyr Jones Parry said.

The report also said it was doubtful that a proper investigation could be carried out with the current Lebanese security apparatus in office.

The pro-Syrian Lebanese government has previously rejected an international inquiry, saying it would cooperate with foreign investigators but it was a matter of national sovereignty not to allow an international probe.

But a statement from the Lebanese president's office late Thursday said Annan spoke to Lahoud and informed him of the broad outlines of the report presented by a U.N. fact-finding team sent to Lebanon after the bombing.

The report said the explosion was caused by a ton of TNT, most likely above the ground.

Members of the opposition praised the report. "It conforms totally with the political vision of the Lebanese opposition," said Ghattas Khoury, a lawmaker from Hariri's parliamentary bloc.

E.6 CBS March 24

A U.N. report into the assassination of former Lebanese Prime Minister Rafik Hariri concluded that Lebanon's investigation into the killing wasn't satisfactory and a new international investigation is needed.

The report, released Thursday, says there was a "distinct lack of commitment" by Lebanese authorities to investigate the crime, and the investigation was not carried out "in accordance with acceptable international standards."

Hariri was killed on Feb. 14 in central Beirut in an explosion that killed 17 other people. The Lebanese opposition has blamed Syria and its Lebanese allies, who have both denied any involvement.

The report does not directly assign blame, saying the causes could not be determined.

"However, it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon," the report said.

President Emile Lahoud urged U.N. Secretary General Kofi Annan late Thursday to do "what is necessary" to uncover the truth behind the Hariri assassination, signaling Lebanon's acceptance of an international inquiry into the Feb. 14 bombing that killed the nation's most prominent politician.

The anti-Syrian Lebanese opposition has accused the government and its Syrian backers of having a hand in the assassination, a charge both the Beirut and Damascus administrations vehemently deny.

The opposition and Hariri's family have insisted on an international investigation, saying they do not trust a Lebanese probe. Some opposition politicians are envisioning an inquiry

with powers to investigate and if necessary bring to justice in an international court those indicted in the case.

It also says Syrian Military intelligence bears responsibility to the extent that it and Lebanese security services failed to provide "security, protection, law and order" in Lebanon.

The opposition and Hariri's family have insisted on an international investigation, saying they have no trust in the Lebanese probe. The report implicitly backed that sentiment, saying the Lebanese investigation "lacks the confidence of the population necessary for its results to be accepted."

Hariri's killing led to turmoil in Lebanon. Mass demonstrations forced the resignation of the Lebanese government and intensified the international campaign for Syria to withdraw its troops from the country.

Syria has now pulled back its troops and intelligence agents into eastern Lebanon toward the border and has been promising to work out their complete removal with the pro-Syrian government in Beirut.

The investigation was carried out by a team led by deputy Irish Police Commissioner Peter Fitzgerald, appointed by U.N. Secretary-General Kofi Annan at the behest of the U.N. Security Council.

In his report, Fitzgerald also faults Syria for interfering in the governing of Lebanon "in a heavy-handed and inflexible manner."

"Without prejudice to the results of the investigation, it is obvious that this atmosphere provided the backdrop for the assassination of Mr. Hariri," he said.

The report went on to demand an international independent commission with the authority to interrogate witnesses, conduct searches and other tasks. Fitzgerald said such an inquiry would be impossible without Lebanon's cooperation.

Later in the day, Lebanese President Emile Lahoud told Annan to do "what is necessary" to unveil the truth.

The report said the explosion was caused by a TNT charge of about 1,000 kilograms, most likely above the ground.

E.7 CBS March 24

A U.N. report into the assassination of former Lebanese Prime Minister Rafik Hariri concluded that Lebanon's investigation into the killing wasn't satisfactory and a new international investigation is needed.

The report, released Thursday, says there was a "distinct lack of commitment" by Lebanese authorities to investigate the crime, and the investigation was not carried out "in accordance with acceptable international standards."

Hariri was killed on Feb. 14 in central Beirut in an explosion that killed 17 other people. The Lebanese opposition has blamed Syria and its Lebanese allies, who have both denied

any involvement.

The report does not directly assign blame, saying the causes could not be determined.

"However, it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon," the report said.

President Emile Lahoud urged U.N. Secretary General Kofi Annan late Thursday to do "what is necessary" to uncover the truth behind the Hariri assassination, signaling Lebanon's acceptance of an international inquiry into the Feb. 14 bombing that killed the nation's most prominent politician.

The anti-Syrian Lebanese opposition has accused the government and its Syrian backers of having a hand in the assassination, a charge both the Beirut and Damascus administrations vehemently deny.

The opposition and Hariri's family have insisted on an international investigation, saying they do not trust a Lebanese probe. Some opposition politicians are envisioning an inquiry with powers to investigate and if necessary bring to justice in an international court those indicted in the case.

It also says Syrian Military intelligence bears responsibility to the extent that it and Lebanese security services failed to provide "security, protection, law and order" in Lebanon.

The opposition and Hariri's family have insisted on an international investigation, saying they have no trust in the Lebanese probe. The report implicitly backed that sentiment, saying the Lebanese investigation "lacks the confidence of the population necessary for its results to be accepted."

Hariri's killing led to turmoil in Lebanon. Mass demonstrations forced the resignation of the Lebanese government and intensified the international campaign for Syria to withdraw its troops from the country.

Syria has now pulled back its troops and intelligence agents into eastern Lebanon toward the border and has been promising to work out their complete removal with the pro-Syrian government in Beirut.

The investigation was carried out by a team led by deputy Irish Police Commissioner Peter Fitzgerald, appointed by U.N. Secretary-General Kofi Annan at the behest of the U.N. Security Council.

In his report, Fitzgerald also faults Syria for interfering in the governing of Lebanon "in a heavy-handed and inflexible manner."

"Without prejudice to the results of the investigation, it is obvious that this atmosphere provided the backdrop for the assassination of Mr. Hariri," he said.

The report went on to demand an international independent commission with the authority to interrogate witnesses, conduct searches and other tasks. Fitzgerald said such an inquiry would be impossible without Lebanon's cooperation.

Later in the day, Lebanese President Emile Lahoud told Annan to do "what is necessary" to unveil the truth.

The report said the explosion was caused by a TNT charge of about 1,000 kilograms, most likely above the ground.

E.8 USA Today March 24

The report, which was obtained by The Associated Press and was to be released later Thursday, says there was a "distinct lack of commitment" by Lebanese authorities to investigate the crime, and the investigation was not carried out "in accordance with acceptable international standards.

In Beirut, Lebanese President Emile Lahoud responded by saying he had told U.N. Secretary-General Kofi Annan to do "what is necessary" to learn who was behind the Feb. 14 killing.

The investigation was carried out by team led by deputy Irish Police Commissioner Peter Fitzgerald, appointed by Annan at the behest of the U.N. Security Council.

In his report, Fitzgerald demanded an "international independent commission" with the authority to interrogate witnesses, conduct searches and other tasks. He said such an inquiry would be impossible without Lebanon's cooperation.

Hariri was killed in central Beirut in an explosion that claimed 17 other lives.

The killing led to political turmoil in Lebanon. Mass demonstrations forced the resignation of the Lebanese government and intensified the international campaign for Syria to withdraw its troops from the country.

Syria has now pulled back its troops and intelligence agents into eastern Lebanon near the border and has been promising to work out their complete removal with the pro-Syrian government in Beirut.

E.9 BBC March 25

Lebanon has indicated it is prepared to co-operate with an international inquiry into last month's killing of former Prime Minister Rafik Hariri.

The move follows a UN report which described Lebanon's own investigation into the bomb attack in Beirut as flawed and inconclusive.

Lebanese authorities criticised the report's findings, saying they were "alien to reality".

And they insisted that any inquiry would have to work with the government.

At a press conference on Friday, Lebanese Foreign Minister Mahmoud Hammoud said the inquiry would be expected to work within an established framework "in co-operation with the state".

"Until now we have not abolished our institutions," he said. "We respect international law and we are committed to the sovereignty of Lebanon."

'Alien to reality'

The UN report did not specify who was behind the 14 February killing but blamed Syria for the political tension that preceded the assassination.

It said the Lebanese inquiry was unsatisfactory.

"The Lebanese investigation process suffers from serious flaws and has neither the capacity nor the commitment to reach a satisfactory and credible conclusion.

"To find the truth it would be necessary to entrust the investigation to an international independent commission."

The report added that Lebanon's security services were unlikely to conduct an adequate inquiry under its current leadership.

Mr Hammoud said the report's conclusions were "alien to reality" adding they were "not based on documents or evidence".

However, Lebanese President Emile Lahoud called on the UN to "do what is necessary" to find Mr Hariri's killers.

But Syria criticised the report, saying it contained "too much rhetoric" and was one-sided.

It comes as Lebanon commemorates 40 days since Mr Hariri's death.

E.10 Seattle Times March 25

A U.N. inquiry into the assassination of former Lebanese Prime Minister Rafik Hariri said yesterday that neighboring Syria bore primary responsibility for the political tension that preceded the attack, and that Syrian President Bashar Assad had threatened Hariri if he opposed extending the term of the country's pro-Syrian president.

The report did not assign blame for the Feb. 14 bomb attack that killed Hariri and 19 others and dramatically increased pressure on Syria to remove its troops and intelligence agents from Lebanon. It called for an international investigation to root out the truth.

The fact-finding mission, led by deputy Irish Police Commissioner Peter Fitzgerald, criticized a Lebanese government investigation as "seriously flawed" and recommended that the leaders of its security services leave office to allow a more credible inquiry. Lebanese investigators removed evidence, including damaged cars, from the scene after the assassination.

President Emile Lahoud urged U.N. Secretary-General Kofi Annan to do whatever is necessary to uncover the truth behind the Hariri assassination.

The U.N. report traced Hariri's growing discontent with Syria's influence in Lebanon and the spiraling power struggle that led to his death.

Hariri had clashed for months with Lahoud over a possible constitutional amendment allowing Lahoud to serve three more years.

The report quoted Hariri's aides and friends as saying Assad told him in a meeting in Damascus last summer that opposing Lahoud was tantamount to opposing the Syrian leader himself.

Assad also reportedly told Hariri that he "would rather break Lebanon over the heads of Hariri and [Druze leader Walid] Jumblatt than see his world in Lebanon broken."

In the weeks after the meeting, Hariri quietly pushed a Security Council resolution, which was adopted last September, demanding a Syrian withdrawal.

The U.N. report emerged against a backdrop of mounting unease in Lebanon. Two car bombs have struck Christian neighborhoods around Beirut in recent days, which sharpened fears that Syria may try to destabilize Lebanon as it withdraws.

The Syrian ambassador to the United Nations, Fayssal Mekdad, called the report "one-sided" and devoid of proof.

E.11 ABC March 25

Lebanese authorities, put on the defensive by a damning U.N. report on security failings, indicated Friday they would accept an international inquiry into the killing of former Prime Minister Rafik al-Hariri.

A U.N. fact-finding team said in a report released on Thursday that Lebanon's own inquiry into Hariri's death on Feb. 14 was seriously flawed and called for an international investigation, a demand long made by the Lebanese opposition.

Lebanon's pro-Syrian officials slammed the report's findings. Opposition figures said it strengthened their calls for such an inquiry and for resignations in the Lebanese security, who they say had a role in Hariri's death.

Hariri's assassination has plunged Lebanon into a political crisis that prompted the resignation of the government, led to mass protests by the opposition and loyalists and piled pressure on Syria to withdraw all its forces from its tiny neighbor.

Asked about the U.N. report's call for an international investigation at a news conference, caretaker Foreign Minister Mahmoud Hammoud said: "We welcome all means that lead us to the truth &hellip We have nothing to hide."

Hammoud said authorities would wait for the U.N. Security Council to issue a resolution on an investigation.

But he criticized the report for blaming Lebanon for not protecting Hariri. "This conclusion is alien to reality &hellip There is no absolute security in any country in the world," he said, adding conclusions were "not based on documents or evidence."

Asked about the report's conclusion that the Lebanese inquiry was flawed, caretaker Justice Minister Adnan Addoum said: "This is a very dangerous accusation that infringes on the dignity of the judicial body and security agencies."

But he admitted Lebanon lacked some inquiry capabilities.

The United States and France, which co-sponsored a resolution calling for Syrian forces to leave, were expected to introduce a resolution in the Security Council calling for an international inquiry, council diplomats said Thursday.

E.12 ABC March 25

U.N. officials issued a report from a fact-finding mission into the assassination of Lebanon's former prime minister Rafik Hariri.

The report did not directly assign blame for the former prime minister's death, but it did say Syria was to blame for the political tensions that existed in Lebanon prior to the attack.

Hariri was killed in a Feb. 14 bombing that killed 17 others on a Beirut seafront street. Many Lebanese blame Syria and its allied Lebanese government for the slaying of Hariri, an opponent of Syrian domination. Syrian officials have agreed to remove 14,000 troops now in Lebanon, but have strongly denied orchestrating the assassination of Hariri.

Following is the text of the executive summary of the U.N. report:

On 14 February 2005, an explosion in downtown Beirut killed twenty persons, among them the former Prime Minister, Rafik Hariri. The United Nations' Secretary-General dispatched a Fact-Finding Mission to Beirut to inquire into the causes, the circumstances and the consequences of this assassination. Since it arrived in Beirut on 25 February, the Mission met with a large number of Lebanese officials and representatives of different political groups, performed a thorough review of the Lebanese investigation and legal proceedings, examined the crime scene and the evidence collected by the local police, collected and analyzed samples from the crime scene, and interviewed some witnesses in relation to the crime.

The specific 'causes' for the assassination of Mr. Hariri cannot be reliably asserted until after the perpetrators of this crime are brought to justice. However, it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon and a failure of the Lebanese State to provide adequate protection for its citizens.

Regarding the circumstances, the Mission is of the view that the explosion was caused by a TNT charge of about 1000 KG placed most likely above the ground. The review of the investigation indicates that there was a distinct lack of commitment on the part of the Lebanese authorities to investigate the crime effectively, and that this investigation was not carried out in accordance with acceptable international standards. The Mission is also of the view that the Lebanese investigation lacks the confidence of the population necessary for its results to be accepted.

The consequences of the assassination could be far-reaching. It seems to have unlocked the gates of political upheavals that were simmering throughout the last year. Accusations and counter-accusations are rife and aggravate the ongoing political polarization. Some accuse the Syrian security services and leadership of assassinating Mr. Hariri because he became an insurmountable obstacle to their influence in Lebanon. Syrian supporters maintain that he was assassinated by "the enemies of Syria"; those who wanted to create

international pressure on the Syrian leadership in order to accelerate the demise of its influence in Lebanon and/or start a chain of reactions that would eventually force a 'regime change' inside Syria itself. Lebanese politicians from different backgrounds expressed to the Mission their fear that Lebanon could be caught in a possible showdown between Syria and the international community, with devastating consequences for Lebanese peace and security.

After gathering the available facts, the Mission concluded that the Lebanese security services and the Syrian Military Intelligence bear the primary responsibility for the lack of security, protection, law and order in Lebanon. The Lebanese security services have demonstrated serious and systematic negligence in carrying out the duties usually performed by a professional national security apparatus. In doing so, they have severely failed to provide the citizens of Lebanon with an acceptable level of security and, therefore, have contributed to the propagation of a culture of intimidation and impunity. The Syrian Military Intelligence shares this responsibility to the extent of its involvement in running the security services in Lebanon.

It is also the Mission's conclusion that the Government of Syria bears primary responsibility for the political tension that preceded the assassination of former Prime Minister Mr. Hariri. The Government of Syria clearly exerted influence that goes beyond the reasonable exercise of cooperative or neighborly relations. It interfered with the details of governance in Lebanon in a heavy-handed and inflexible manner that was the primary reason for the political polarization that ensued. Without prejudice to the results of the investigation, it is obvious that this atmosphere provided the backdrop for the assassination of Mr. Hariri.

It became clear to the Mission that the Lebanese investigation process suffers from serious flaws and has neither the capacity nor the commitment to reach a satisfactory and credible conclusion. To find the truth, it would be necessary to entrust the investigation to an international independent commission, comprising the different fields of expertise that are usually involved in carrying out similarly large investigations in national systems, with the necessary executive authority to carry out interrogations, searches, and other relevant tasks. Furthermore, it is more than doubtful that such an international commission could carry out its tasks satisfactorily - and receives the necessary active cooperation from local authorities - while the current leadership of the Lebanese security services remains in office.

It is the Mission's conclusion that the restoration of the integrity and credibility of the Lebanese security apparatus is of vital importance to the security and stability of the country. A sustained effort to restructure, reform and retrain the Lebanese security services will be necessary to achieve this end, and will certainly require assistance and active engagement on the part of the international community.

Finally, it is the Mission's view that international and regional political support will be necessary to safeguard Lebanon's national unity and to shield its fragile polity from unwarranted pressure. Improving the prospects of peace and security in the region would

offer a more solid ground for restoring normalcy in Lebanon.

E.13 Boston Globe March 25

President Bashar Assad of Syria threatened former Lebanese prime minister Rafik Hariri with "physical harm" last summer if Hariri challenged Assad's dominance over Lebanese political life, contributing to a climate of violence that led to the Feb. 14 slayings of Hariri and 19 others, according to testimony in a report released yesterday by a UN fact-finding team.

The report, which calls for an international investigation into Hariri's slaying, describes an August meeting in Damascus at which Assad ordered the Lebanese billionaire to support amending the Lebanon Constitution, according to testimony from "various" sources who discussed the meeting with Hariri. The amendment, approved Sept. 3, allowed Emile Lahoud, the Syrian-backed Lebanese president, to remain in office for three more years.

Assad said that "Lahoud should be viewed as his personal representative" in Lebanon and that "opposing him is tantamount to opposing Assad himself," the report states. Assad then warned that he "would rather break Lebanon over the heads of" Hariri and influential Druze political leader Walid Jumblatt "than see his word in Lebanon broken."

The UN team, which was headed by Ireland's deputy police commissioner, Peter FitzGerald, alleged that Syrian-controlled Lebanese authorities exhibited a "distinct lack of commitment" to conducting a credible investigation into Hariri's assassination by tampering with evidence and failing to pursue promising leads.

FitzGerald stopped short of accusing Syria and its Lebanese allies of detonating the bomb that killed Assad's major political rival in Lebanon. But he contended that Syria "bears primary responsibility for the political tension that preceded" Hariri's assassination.

In the report, FitzGerald said the international investigative team "would need executive authority to carry out its interrogations, searches, and other relevant tasks." But he added that it was "more than doubtful" that an international investigation into the crime could succeed as long as the current leadership in Lebanon's Syrian-backed security establishment remains in power.

The 20-page report, presented to the UN Security Council this afternoon by Secretary General Kofi Annan, represents the most damning official account of Syria's role in Lebanon.

Syria's UN envoy, Fayssal Mekdad, questioned the need for an international investigation into the killing, saying Lebanese officials were capable of doing the job. Mekdad also denied that Assad and other Syrian authorities had played any role in Hariri's death. "I assure you we don't deal this way," Mekdad said.

FitzGerald said that a "single individual or small terrorist group" lacked the capacity to carry out such a sophisticated attack, which required "considerable finance, military precision in its execution, [and] substantial logistical support."

E.14 Boston Globe March 25

Lebanon's pro-Syrian leaders Friday rejected a sharply critical U.N. report blaming Damascus for stoking tensions that may have led to the assassination of former premier Rafik Hariri, but they grudgingly accepted a U.N. investigation into the slaying.

The report embarrassed the Lebanese government and its backers, saying Syria's president personally threatened Hariri with harm for his opposition to Damascus' domination and criticizing Lebanon for a halfhearted investigation into who killed him in a Feb. 14 bombing.

But a U.N. inquiry could be more damaging, if more aggressive investigators have powers to subpoena officials and carry out searches as the report recommends. The report also said a probe would be difficult while Lebanon's security chiefs are in place, bolstering a top demand of Lebanon's anti-Syrian opposition.

It was unclear if the government would consent to broad powers for an investigation. Asked Friday how Lebanon's justice system would deal with an inquiry if established by the U.N. Security Council, Lebanese Justice Minister Adnan Addoum had a list of technical and legal concerns.

"This is a subject that requires a long research in international law, the limits and jurisdictions and what kind of court is right for this subject," he said, though he insisted Lebanon will abide by international decisions.

But, under stepped-up pressure, the government may have a hard time avoiding an aggressive inquiry. Told by U.N. chief Kofi Annan about the report's contents, Lebanese President Emile Lahoud late Thursday agreed to a U.N. investigation, which his government had repeatedly rejected.

Foreign Minister Mahmoud Hammoud said Friday the government "welcomes all means" to find the truth about the bombing that killed Hariri and 17 others on a Beirut seafront street.

Hariri's death sparked massive demonstrations in Lebanon that shook Syria's domination of its politics, disrupting the government and helping force Damascus to pull back its 14,000 troops in Lebanon. Many Lebanese accuse the governments in Beirut and Damascus of being behind the slaying, a claim both vehemently deny.

Discussing a U.N. inquiry, Chibli Mallat, an international law professor at St. Joseph's University in Beirut, told LBC television that "under international law, immunity no longer applies to anyone, including the presidents of Lebanon or Syria, if they are asked to testify or in case they were indicted."

Government ministers were defensive Friday as they held news conferences to dispute the report's findings that their government had botched, if not outright manipulated, its own investigation into Hariri's killing.

Defense Minister Abdul-Rahim Murad was visibly angry over claims Syria and Lebanon

were lax in security. "Why American forces in Iraq do not uncover the car bombs and the operations against civilians and the American army?" he asked at a news conference.

Hammoud said the U.N. mission overstepped its mandate in accusing the government of negligence. "This descriptive report ... reaches conclusions that are not based on documented evidence," Hammoud said.

Continued...

E.15 CNN March 25

A fact-finding team investigating last month's assassination of former Lebanese Prime Minister Rafik Hariri has blamed Syria's government for the political tension that preceded the killing, according to a U.N. report released Thursday.

The government of Syria "interfered" with governance in Lebanon in a heavy-handed way that was "the primary reason for the political polarization that ensued."

"It is obvious that this atmosphere provided the backdrop for the assassination of Mr. Hariri," the report says.

The investigative team was assembled by U.N. Secretary-General Kofi Annan to look into "the causes, circumstances and consequences of the assassination," which resulted in large-scale demonstrations against Syria's troop presence in Lebanon and the resignation of Prime Minister Omar Karami's pro-Syrian government.

The massive bomb blast along Beirut's waterfront killed 20 people and wounded more than 100.

A spokesman for Lebanese President Emile Lahoud said Annan called the president and outlined the report's contents. Lahoud urged Annan "to take appropriate measures to unveil the truth as soon as possible," the spokesman said.

Lebanese parliament member Ghattas Khoury, a member of Hariri's party, said he agreed with the report's conclusions.

"I think the report is fair. It coincides with the view of the opposition that security agencies were completely negligent. ... up to the level of direct involvement or even probably conspiracy," he told CNN.

"There is a need to have foreign international investigation ... to clear up the truth about who was behind the assassination," he added.

Hariri was the chief opposition figure in Lebanon to push for the exit of Syrian troops and intelligence officers from his country following last year's passage of U.N. Resolution 1559, which called for a full withdrawal.

Last week, Syria began moving its 14,000 troops to the Bekaa Valley near the border with Lebanon and promised to bring all the troops and intelligence officials across the border into Syria as soon as possible.

According to the report, the specific causes for the assassination will not be known until after the killers are brought to justice.

"However, it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon and a failure of the Lebanese state to provide adequate protection for its citizens," investigators concluded.

The report calls Lebanon's security services "negligent," and accuses them of contributing to the "propagation of a culture of intimidation and impunity."

Protection must be beefed up to boost the nation's security and credibility, it says.

E.16 CNN March 25

A fact-finding team investigating last month's assassination of former Lebanese Prime Minister Rafik Hariri has blamed Syria's government for the political tension that preceded the killing, according to a U.N. report released Thursday.

The government of Syria "interfered" with governance in Lebanon in a heavy-handed way that was "the primary reason for the political polarization that ensued."

"It is obvious that this atmosphere provided the backdrop for the assassination of Mr. Hariri," the report says.

The investigative team was assembled by U.N. Secretary-General Kofi Annan to look into "the causes, circumstances and consequences of the assassination," which resulted in large-scale demonstrations against Syria's troop presence in Lebanon and the resignation of Prime Minister Omar Karami's pro-Syrian government.

The massive bomb blast along Beirut's waterfront killed 20 people and wounded more than 100.

A spokesman for Lebanese President Emile Lahoud said Annan called the president and outlined the report's contents. Lahoud urged Annan "to take appropriate measures to unveil the truth as soon as possible," the spokesman said.

Lebanese parliament member Ghattas Khoury, a member of Hariri's party, said he agreed with the report's conclusions.

"I think the report is fair. It coincides with the view of the opposition that security agencies were completely negligent. ... up to the level of direct involvement or even probably conspiracy," he told CNN.

"There is a need to have foreign international investigation ... to clear up the truth about who was behind the assassination," he added.

Hariri was the chief opposition figure in Lebanon to push for the exit of Syrian troops and intelligence officers from his country following last year's passage of U.N. Resolution 1559, which called for a full withdrawal.

Last week, Syria began moving its 14,000 troops to the Bekaa Valley near the border with Lebanon and promised to bring all the troops and intelligence officials across the border

into Syria as soon as possible.

According to the report, the specific causes for the assassination will not be known until after the killers are brought to justice.

"However, it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon and a failure of the Lebanese state to provide adequate protection for its citizens," investigators concluded.

The report calls Lebanon's security services "negligent," and accuses them of contributing to the "propagation of a culture of intimidation and impunity."

Protection must beefed up to boost the nation's security and credibility, it says.

E.17 New York Times March 25

A toughly worded United Nations report into the assassination of the former Lebanese prime minister, Rafik Hariri, concluded Thursday that heavy-handed Syrian interference in Lebanese affairs had created the polarizing tensions that led to his death and that a deeply flawed local investigation had obstructed efforts to find his killers.

At one point, the report said, quoting aides to Mr. Hariri, Syria's leader personally threatened Mr. Hariri.

The author, Patrick Fitzgerald, a deputy police commissioner of Ireland, called for an investigation by an independent commission as the only way of uncovering the truth behind Mr. Hariri's killing.

The Syrian-backed Lebanese government previously rejected such an inquiry, saying it would violate national sovereignty, but Lebanon's president, Emile Lahoud, issued a statement in Beirut on Thursday night after speaking with Secretary General Kofi Annan by telephone that appeared to give approval for an international investigation.

"President Lahoud asked the secretary general to do what's necessary to reveal the truth in the crime as soon as possible," the statement said.

In a covering letter to the Security Council, Mr. Annan endorsed the call for an international investigation, saying the report raised "very serious and troubling allegations."

Mr. Hariri was killed on Feb. 14 when his motorcade was bombed in central Beirut in a blast that killed 19 other people.

The assassination prompted mass demonstrations, the resignation of the Lebanese government and accusations against Syria from the Lebanese opposition, and it intensified the withdrawal of Syrian troops and intelligence agents that Damascus now promises to complete by May.

Meeting a day after the assassination, the Security Council dispatched the fact-finding mission to "report urgently on the circumstances, causes and consequences of the assassination."

While it said it could not assign direct blame for the killing, the mission's 19-page

report said the government of Syria "bears primary responsibility for the political tension that preceded the assassination." It said Syria's interference in Lebanon was "heavy-handed and inflexible" which, combined with inept Lebanese security, was responsible for "political polarization" that "provided the backdrop" for the assassination.

The mission said it had been told by a number of people close to Mr. Hariri that he had reported that in his last meeting with President Bashar al-Assad of Syria, the Syrian leader had threatened him with physical harm if he continued his campaign to assert Lebanese independence from Syria. The report said the Syrians had refused to discuss the meeting with the mission's investigators.

"It is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon and a failure of the Lebanese state to provide adequate protection for its citizens," it said.

It accused the Lebanese security services of "serious and systematic negligence" and implied that the authorities had obstructed the team's work. "There was a distinct lack of commitment on the part of the Lebanese authorities to investigate the crime effectively," the report said. The Lebanese-Syrian security combine, it said, bred in Lebanon "a culture of intimidation and impunity."

It said too that Lebanese investigators inspired no trust in the Lebanese people and asserted that the current leadership of the Lebanese security services would have to be replaced before any credible international investigation could take place.

At the United Nations, Fayssal Mekdad, the Syrian ambassador, was dismissive of the report, saying Mr. Fitzgerald had spent too much time talking to opponents of Syria. "It seems to me that he deals only with the opposition and those who want to accuse Syria of something," he said. "He should have been more objective in analyzing the overall situation."

E.18 Washington Post March 25

Syrian President Bashar Assad threatened former Lebanese prime minister Rafiq Hariri with "physical harm" last summer if Hariri challenged Assad's dominance over Lebanese political life, contributing to a climate of violence that led to the Feb. 14 slayings of Hariri and 19 others, according to testimony in a report released Thursday by a U.N. fact-finding team.

The report, which calls for an international investigation into Hariri's death, describes an August meeting in Damascus at which Assad ordered the Lebanese billionaire to support amending Lebanon's constitution, according to testimony from "various" sources who discussed the meeting with Hariri. The amendment, approved Sept. 3, allowed Emile Lahoud, the Syrian-backed Lebanese president, to remain in office for three more years.

Assad said that "Lahoud should be viewed as his personal representative" in Lebanon

and that "opposing him is tantamount to opposing Assad himself," the report states. Assad then warned that he "would rather break Lebanon over the heads of" Hariri and influential Druze political leader Walid Jumblatt "than see his word in Lebanon broken."

The U.N. team, which was headed by Ireland's deputy police commissioner, Peter FitzGerald, charged that Syrian-controlled Lebanese authorities exhibited a "distinct lack of commitment" to conducting a credible investigation into Hariri's assassination by tampering with evidence and failing to pursue promising leads.

FitzGerald stopped short of accusing Syria and its Lebanese allies of detonating the 2,200-pound bomb that killed Assad's major political rival in Lebanon. But he charged that Syria "bears primary responsibility for the political tension that preceded" Hariri's assassination.

In the report, FitzGerald said that the international investigative team "would need executive authority to carry out its interrogations, searches and other relevant tasks." But he added that it was "more than doubtful" that an international investigation into the crime could succeed as long as the leadership in Lebanon's Syrian-backed security establishment remains in power.

The 20-page report, presented to the U.N. Security Council on Thursday afternoon by Secretary General Kofi Annan, represents the most damning official account of Syria's role in Lebanon. Annan phoned Assad and Lahoud to warn them of the report's findings.

Syria's U.N. envoy, Fayssal Mekdad, questioned the need for an international investigation into the killing, saying that Lebanese officials were capable of doing the job. Mekdad also denied that Assad and other Syrian authorities had played any role in Hariri's death. "I assure you we don't deal this way," Mekdad told reporters.

The crisis in Lebanon has unfolded against mounting political resistance to Syrian dominance in the former French colony. Syria, which first sent troops into Lebanon in 1975, has long feared a strong Lebanese opposition movement. Members of Assad's ruling Baath Party have economic and political interests in Lebanon, and Hariri consistently voiced opposition to Syrian influence in Lebanon.

Syria's drive to amend Lebanon's constitution last year fueled stiff opposition from Hariri and other opposition figures seeking to gain power. Hariri and his party ultimately supported the extension of Lahoud's presidency. But the popular Lebanese politician and businessman subsequently resigned in protest and played a role in securing support from French President Jacques Chirac, a personal friend, and the United States to adopt a Sept. 2 Security Council resolution demanding that Syria withdraw its 20,000 troops and intelligence agents from Lebanon.

"The Syrian leadership held Hariri personally responsible for the adoption of the resolution," FitzGerald said sources told him. "Clearly, Mr. Hariri's assassination took place on the backdrop of his power struggle with Syria."

The report says that Hariri's government security detail, which included 40 agents, was

reduced to eight after he resigned as prime minister, despite continued threats against his life. "The Lebanese security apparatus failed to provide proper protection for Mr. Hariri," FitzGerald said.

The report says that Hariri was killed as his convoy passed over a massive explosive placed on the road outside the Hotel Saint-Georges in central Beirut.

Following the attack, Lebanese authorities failed to properly secure the site and cleared it of key evidence, including the six vehicles in Hariri's convoy, according to the report.

The police failed to shut down a broken water main that flooded the crime scene, washing away important evidence.

"Important evidence was either removed or destroyed without record," FitzGerald said.

The report also charges that Lebanese investigators neglected to trace a "suspect" white pickup truck that slowed down at the crime scene in the minutes before the explosion. Nor did they interview potential witnesses, a failure that amounted to "gross negligence."

FitzGerald cast doubt over reports that an alleged Lebanese militant, Ahmad Abu Adas, 22, who claimed responsibility for the assassination, carried it out. Adas described himself as a member of a previously unknown militant group, Nasra and Jihad Group in Greater Syria, in a videotaped confession broadcast by the Arab language network Al-Jazeera.

FitzGerald said that a "single individual or small terrorist group" lacked the capacity to carry out such an attack, which required "considerable finance, military precision in its execution, [and] substantial logistical support."

E.19 Washington Post March 25

Syrian President Bashar Assad threatened former Lebanese prime minister Rafiq Hariri with "physical harm" last summer if Hariri challenged Assad's dominance over Lebanese political life, contributing to a climate of violence that led to the Feb. 14 slayings of Hariri and 19 others, according to testimony in a report released Thursday by a U.N. fact-finding team.

The report, which calls for an international investigation into Hariri's death, describes an August meeting in Damascus at which Assad ordered the Lebanese billionaire to support amending Lebanon's constitution, according to testimony from "various" sources who discussed the meeting with Hariri. The amendment, approved Sept. 3, allowed Emile Lahoud, the Syrian-backed Lebanese president, to remain in office for three more years.

Assad said that "Lahoud should be viewed as his personal representative" in Lebanon and that "opposing him is tantamount to opposing Assad himself," the report states. Assad then warned that he "would rather break Lebanon over the heads of" Hariri and influential Druze political leader Walid Jumblatt "than see his word in Lebanon broken."

The U.N. team, which was headed by Ireland's deputy police commissioner, Peter FitzGerald, charged that Syrian-controlled Lebanese authorities exhibited a "distinct lack of

commitment" to conducting a credible investigation into Hariri's assassination by tampering with evidence and failing to pursue promising leads.

FitzGerald stopped short of accusing Syria and its Lebanese allies of detonating the 2,200-pound bomb that killed Assad's major political rival in Lebanon. But he charged that Syria "bears primary responsibility for the political tension that preceded" Hariri's assassination.

In the report, FitzGerald said that the international investigative team "would need executive authority to carry out its interrogations, searches and other relevant tasks." But he added that it was "more than doubtful" that an international investigation into the crime could succeed as long as the leadership in Lebanon's Syrian-backed security establishment remains in power.

The 20-page report, presented to the U.N. Security Council on Thursday afternoon by Secretary General Kofi Annan, represents the most damning official account of Syria's role in Lebanon. Annan phoned Assad and Lahoud to warn them of the report's findings.

Syria's U.N. envoy, Fayssal Mekdad, questioned the need for an international investigation into the killing, saying that Lebanese officials were capable of doing the job. Mekdad also denied that Assad and other Syrian authorities had played any role in Hariri's death. "I assure you we don't deal this way," Mekdad told reporters.

The crisis in Lebanon has unfolded against mounting political resistance to Syrian dominance in the former French colony. Syria, which first sent troops into Lebanon in 1975, has long feared a strong Lebanese opposition movement. Members of Assad's ruling Baath Party have economic and political interests in Lebanon, and Hariri consistently voiced opposition to Syrian influence in Lebanon.

Syria's drive to amend Lebanon's constitution last year fueled stiff opposition from Hariri and other opposition figures seeking to gain power. Hariri and his party ultimately supported the extension of Lahoud's presidency. But the popular Lebanese politician and businessman subsequently resigned in protest and played a role in securing support from French President Jacques Chirac, a personal friend, and the United States to adopt a Sept. 2 Security Council resolution demanding that Syria withdraw its 20,000 troops and intelligence agents from Lebanon.

"The Syrian leadership held Hariri personally responsible for the adoption of the resolution," FitzGerald said sources told him. "Clearly, Mr. Hariri's assassination took place on the backdrop of his power struggle with Syria."

The report says that Hariri's government security detail, which included 40 agents, was reduced to eight after he resigned as prime minister, despite continued threats against his life. "The Lebanese security apparatus failed to provide proper protection for Mr. Hariri," FitzGerald said.

The report says that Hariri was killed as his convoy passed over a massive explosive placed on the road outside the Hotel Saint-Georges in central Beirut.

Following the attack, Lebanese authorities failed to properly secure the site and cleared it of key evidence, including the six vehicles in Hariri's convoy, according to the report.

The police failed to shut down a broken water main that flooded the crime scene, washing away important evidence.

"Important evidence was either removed or destroyed without record," FitzGerald said.

The report also charges that Lebanese investigators neglected to trace a "suspect" white pickup truck that slowed down at the crime scene in the minutes before the explosion. Nor did they interview potential witnesses, a failure that amounted to "gross negligence."

FitzGerald cast doubt over reports that an alleged Lebanese militant, Ahmad Abu Adas, 22, who claimed responsibility for the assassination, carried it out. Adas described himself as a member of a previously unknown militant group, Nasra and Jihad Group in Greater Syria, in a videotaped confession broadcast by the Arab language network Al-Jazeera.

FitzGerald said that a "single individual or small terrorist group" lacked the capacity to carry out such an attack, which required "considerable finance, military precision in its execution, [and] substantial logistical support."

E.20 MSNBC March 26

BEIRUT, Lebanon - A bomb set off a raging inferno in an industrial area of a mainly Christian neighborhood of Beirut on Saturday, injuring at least three foreign workers.

Antoine Gebara, mayor of the northeastern Beirut area of Bouchrieh, said the explosion was caused by a bomb placed near the buildings in an industrial area.

"It appears it is an explosive charge that was placed there," Gebara told Lebanese Broadcasting Corp.

"They must love us - we got it twice in a week," he said referring to an explosion in the nearby predominantly Christian neighborhood of Jdeideh last Saturday that injured nine people. A bomb on Wednesday killed three people in a Christian commercial center.

Blast preceded Mass

Witnesses said the blast on the eve of the Easter holiday occurred three hours before Catholics were to head to a midnight Mass.

An Associated Press photographer at the scene said at least six buildings were ablaze in the Bouchrieh industrial zone, one of Beirut's largest. Firefighters battled desperately to contain huge orange flames leaping from blackened factory windows.

A military officer said at least three Asian workers were wounded. He said it was difficult to know the nature of the explosion because of the fire. He said some of the factories in the area contained highly flammable material.

An explosives expert said Saturday's bomb, which was placed between a car and a building containing a wood-working factory, weighed about 55 pounds.

Civil Defense officers and the Red Cross were calling on people to stay away from the area, fearing the spread of fire and more explosions caused by flammable materials and fuels in the factories.

Political turmoil

Lebanon has been tense since the Feb. 14 assassination of former Prime Minister Rafik Hariri and the subsequent withdrawal of Syrian troops to the east of the country and Syria.

The three bombings since March 19 have targeted Christian, anti-Syrian strongholds in Lebanon, killing a total of three people and injuring 16, causing fears of the return of the sectarian violence that plagued Lebanon during the 1975-90 civil war.

The motive behind the attacks wasn't immediately clear, but Lebanese opposition leaders have blamed Syrian security agents and pro-Damascus Lebanese authorities, saying they wanted to sow fear in the community.

Opposition leader Walid Jumblatt held pro-Syrian Lebanese security agents responsible, saying they were trying to intimidate people. The Druse leader said he expected more car bombs in the coming days and in the run-up to parliamentary elections scheduled to be held by May.

Another opposition leader, Butros Harb, said the explosions were "a political message from the authorities and those behind them" aimed at "terrorizing" the Lebanese people who are demanding freedom and sovereignty.

Harb told Al-Arabiya TV that Lebanese will not be cowed by such acts and will continue seeking independence.

The explosion sent up a black column of smoke and panicked residents ran into the street while civil defense workers rushed to the area to extinguish the blaze.

E.21 Boston Globe March 26

Senior Lebanese officials yesterday rejected a UN report blaming Syria for tensions that led to the slaying of former prime minister Rafik Hariri, saying the UN mission exceeded its authority in accusing the government of negligence.

The report from a UN fact-finding mission was sharply critical of Syria and its allied Lebanese government. It said that there was evidence Syria's president threatened Hariri with physical harm and that the Beirut government showed a lack of commitment to finding out who killed him, bungling and outright manipulating the investigation.

Many Lebanese blame Syria and the Lebanese government for the slaying of Hariri – an opponent of Syrian domination – in a Feb. 14 bombing that also killed 17 others on a seafront street in Beirut. Officials in Damascus and Beirut vehemently deny any role in the killing.

The report stopped short of blaming Syria in the killing, but did say it was to blame for the political tensions in the country before Hariri's death.

Mahmoud Hammoud, Lebanon's foreign minister, said the UN fact-finding team, which released its report Thursday, had gone beyond its mandate.

"The [UN] mission had no authority to allow it to reach these conclusions," he said. "We see this as infringement on the role of the Lebanese government."

Still, he insisted that the government "welcomes all means" to find the truth about the bombing.

Late Thursday, President Emile Lahoud of Lebanon urged UN Secretary General Kofi Annan to do "what is necessary" to uncover the truth behind Hariri's assassination, signaling Lebanon's acceptance of an international inquiry that it had been persistently rejecting since the bombing.

The UN report does not directly assign blame for Hariri's death. But it did say that "it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon."

Hammoud rejected this, saying tensions were caused by the United Nations' call for Syria to withdraw its troops from Lebanon in Resolution 1559.

"We say that tension began when signals starting coming from abroad that a resolution was to be issued by the Security Council – it later became 1559. This resolution pushed the atmosphere toward political polarization," Hammoud said.

Justice Minister Adnan Addoum emphasized that the UN report was not a legal opinion and rejected claims of evidence tampering.

"We consider it a technical security document and it cannot be considered a legal and judicial document," he told the news conference.

Still, Interior Minister Suleiman Franjieh acknowledged "flaws" in the security system, as the report noted.

The report said there was a "distinct lack of commitment" by the authorities to investigate the crime, and it detailed a host of flaws, including the disappearance of crucial evidence and tampering at the scene of the blast. Parts of a pickup were brought to the scene, placed in the crater, and photographed as evidence, it said.

Syrian military intelligence shares responsibility with Lebanese security forces for not providing "security, protection, law, and order" in Lebanon, the report said.

"This is far from reality," Hammoud said.

Addoum denied that car parts were put in the crater. "The proof is that the wreckage of the car was found in the sea near the site and was retrieved by divers from the international experts," he said.

E.22 USA Today March 26

The latest attack, targeting an industrial area in Beirut's northeastern Bouchrieh area, raised tensions another notch in Lebanon, which has been gripped by political turmoil over

Syria's presence since the Feb. 14 assassination of former premier Rafik Hariri.

A 55 pound bomb was placed between a car and a furniture factory, said Lebanon's police chief, Maj. Gen. Sarkis Tadros, citing an explosives expert. The blast destroyed nearby cars, shattered windows and left a crater that was 3 feet deep and 10 feet wide.

A Lebanese woman and two Indian workers were injured, as were two civil defense workers working on extinguishing the fire that engulfed at least six buildings, security officials said.

"They must love us - we got it twice in a week," Bouchrieh mayor Antoine Gebara told Lebanese Broadcasting Corp. He was referring to last Saturday's explosion in the nearby predominantly Christian neighborhood of Jdeideh that injured nine people. Five days later, another bomb blast killed three people near the port city of Jounieh, Lebanon's Christian heartland.

Witnesses said the blast on the eve of the Easter holiday occurred three hours before Catholics were to head to a midnight Mass.

The motive behind the latest attacks wasn't clear, but Lebanese opposition leaders have blamed Syrian security agents and pro-Damascus Lebanese authorities for trying to show a need for Syria's military presence in Lebanon in the midst of a Syrian troop withdrawal.

Each attack has targeted Christian, anti-Syrian strongholds, raising fears of the return of the sectarian violence that plagued Lebanon during the 1975-90 civil war.

"They (Syrians) think they can destroy Lebanese national unity this way. But the Lebanese will remain steadfast till infinity," exiled Christian opposition leader Michel Aoun told Al-Arabiya TV.

Aoun said the situation calls for "changing the security organizations related to Syria. This can't be delayed."

The death of Hariri, who opposed Syria's presence, sparked massive demonstrations in Lebanon that disrupted the government and helped force Damascus to pull back its 14,000 troops to eastern Lebanon under international pressure. Many Lebanese accuse the governments in Beirut and Damascus of being behind the slaying, a claim both vehemently deny.

About 1,000 of the 10,000 Syrian soldiers remaining in eastern Lebanon's Bekaa Valley had started heading home in recent days, a Lebanese military official said Saturday. The redeployments follow the return to Syria of 4,000 soldiers in the first phase of the troop withdrawal that was completed March 17.

Lebanon's pro-Syrian Defense Minister Abdul-Rahim Murad warned that the Lebanese army may not be able to handle security if Syrian forces leave the eastern Bekaa Valley, a strategically important region for Syria's own security, particularly in facing rival Israel.

The Bekaa, which covers 45% of Lebanese territory, "needs a lot of military forces," Murad told reporters Friday, hinting that Syrian troops may still be needed in Lebanon.

Murad, who hails from the Bekaa, said the U.S. ambassador asked Lebanon's army com-

mander recently about Lebanese army readiness to replace Syrian forces in eastern Lebanon. Murad said the commander replied that "the conditions of the military establishment do not permit this new role in the Bekaa because numerically the army is not enough."

Lebanese opposition leader Walid Jumblatt rejected Murad's comments and renewed calls on Lebanese security chiefs to resign in the wake of a U.N. report this week that criticized Syria and its allied Lebanese government in connection with Hariri's killing.

The report also recommended an international investigation into Hariri's murder, but added such a probe would be difficult while Lebanon's security chiefs are in place.

"It is not possible to carry out a just, clear and transparent investigation if the heads of (security) agencies remained in place," Jumblatt said Saturday. Legislator Bahiya Hariri, the slain leader's sister, also demanded the resignations.

Jumblatt said he expected more car bombs in the coming days and in the run-up to parliamentary elections scheduled to be held by May.

The pro-Syrian camp, however, accused opposition forces of seeking the instability to invite international intervention in Lebanon.

"I think what is going on is an attempt to internationalize the Lebanese situation to allow for sending troops to Lebanon," said Karim Pakradouni, leader of the pro-government Christian Phalange party, adding he did not believe security agencies were to blame.

Syrian soldiers have been based in Lebanon since 1976, when they arrived ostensibly to provide a stabilizing force in the war-torn country. They remained after the end of hostilities, controlling all important political and security issues in Lebanon.

E.23 Washington Post March 26

The Bush administration is reaching out to the Syrian opposition because of growing concerns that unrest in Lebanon could spill over and suddenly destabilize Syria, which borders four countries pivotal to U.S. Middle East policy – Israel, Iraq, Lebanon and Turkey, U.S. and Syrian sources said.

In an interview, Secretary of State Condoleezza Rice said yesterday that the United States is talking to "as many people as we possibly can" about the situation in Syria, as well as in Lebanon, to ensure that Washington is prepared in the event of yet another abrupt political upheaval.

"What we're trying to do is to assess the situation so that nobody is blindsided, because events are moving so fast and in such unpredictable directions that it is only prudent at this point to know what's going on," Rice told Washington Post editors and reporters, citing "the possibility for what I often call discontinuous events, meaning that you were expecting them to go along like this and all of a sudden they go off in this direction, in periods of change like this. So we're going to look at all the possibilities and talk to as many people as we possibly can."

The Thursday meeting, hosted by new State Department "democracy czar" Elizabeth Cheney, brought together senior administration officials from Vice President Cheney's office, the National Security Council and the Pentagon and about a dozen prominent Syrian Americans, including political activists, community leaders, academics and an opposition group, a senior State Department official said.

The opposition group comes from the Syria Reform Party, a small U.S.-based Syrian organization often compared to the Iraqi National Congress led by former exile Ahmed Chalabi. The INC, which led the campaign to oust former Iraqi president Saddam Hussein, had widespread U.S. financial and political support from both the Clinton and Bush administrations, as well as Congress.

U.S. officials, however, yesterday denied that the meeting was intended to coordinate efforts to oust Syrian President Bashar Assad's government.

"That would be a monumental distortion," a senior State Department official said. "But it was a discussion about supporting reform and change in the region and specifically Syria – and how we can help that and work with people in the region and Syria to support that process."

The U.S. outreach is a direct result of President Bush's discussion last month with French President Jacques Chirac, said U.S and European officials. Advising against any discussion of "regime change," Chirac told Bush that the Damascus government was unlikely to survive the withdrawal of Syrian forces from Lebanon. The French president predicted that free elections in Lebanon would in turn force change inside Syria, possibly unraveling Assad's government, U.S. sources said.

Since that Feb. 21 meeting, the Bush administration has begun looking at possible political options in Syria, said analysts familiar with the U.S. thinking. "They're taking seriously that a consequence of getting out of Lebanon will be the collapse of the Assad regime, and they're looking around for alternatives," said Flynt Leverett, former senior director for Middle East affairs at the National Security Council under Bush.

The Syrian Americans who attended the meeting urged the administration to take tentative steps to pressure Damascus, such as having Bush call for greater freedoms and release of political prisoners, said Farid Ghadry, president of the Syrian Reform Party.

The delegation also sought support for lawsuits in U.S. courts against Syrian officials engaged in human rights abuses, an option available under the Alien Tort Claims Act, Ghadry said. The 1789 law grants jurisdiction to U.S. federal courts over "any civil action by an alien for a tort only, committed in violation of the law of nations or a treaty of the United States."

Ghadry said the Syrian opposition was encouraged by the "open and constructive" meeting, which was attended by key players in the administration's democracy policy such as John Hannah from Cheney's office, Robert Danin from the National Security Council and the Pentagon's David Schenker.

"They wanted to hear from us how they can help in extending the message of freedom and democracy in Syria," said Ghadry, who left his homeland 30 years ago, when he was 10, and formed his party after the Sept. 11, 2001, terrorist attacks. "They listened and took a lot of notes. We felt from the responses that they understand these are important issues."

Some U.S. analysts and other Syrian Americans warned that the Syrian Reform Party and its allies are unrepresentative and too small to have any impact.

"Its membership is extremely thin and is not taken seriously. It's almost unheard-of in Syria," said Murhaf Jouejati, director of George Washington University's Middle East Studies Program.

On Lebanon, Rice said the United States is waiting to hear recommendations from U.N. envoy Terje Roed-Larsen on how to support spring elections there. "The main thing is just to help the Lebanese opposition and others, the entire Lebanese political space [and] people to get organized so that they can have a competitive, free and fair election," she said.

"I would suspect that if the U.N. comes back and says [do election] monitoring, people will be very supportive of that," Rice added. "Perhaps if there's need for nongovernmental organizations to do training or the kind of things that have been done in other places, I'm quite sure that people would be prepared to do that."

E.24 Washington Post March 26

The Bush administration is reaching out to the Syrian opposition because of growing concerns that unrest in Lebanon could spill over and suddenly destabilize Syria, which borders four countries pivotal to U.S. Middle East policy – Israel, Iraq, Lebanon and Turkey, U.S. and Syrian sources said.

In an interview, Secretary of State Condoleezza Rice said yesterday that the United States is talking to "as many people as we possibly can" about the situation in Syria, as well as in Lebanon, to ensure that Washington is prepared in the event of yet another abrupt political upheaval.

"What we're trying to do is to assess the situation so that nobody is blindsided, because events are moving so fast and in such unpredictable directions that it is only prudent at this point to know what's going on," Rice told Washington Post editors and reporters, citing "the possibility for what I often call discontinuous events, meaning that you were expecting them to go along like this and all of a sudden they go off in this direction, in periods of change like this. So we're going to look at all the possibilities and talk to as many people as we possibly can."

A meeting Thursday, hosted by new State Department "democracy czar" Elizabeth Cheney, brought together senior administration officials from Vice President Cheney's office, the National Security Council and the Pentagon and about a dozen prominent Syrian Amer-

icans, including political activists, community leaders, academics and an opposition group, a senior State Department official said.

The opposition group comes from the Syria Reform Party, a small U.S.-based Syrian organization often compared to the Iraqi National Congress led by former exile Ahmed Chalabi. The INC, which led the campaign to oust former Iraqi president Saddam Hussein, had widespread U.S. financial and political support from both the Clinton and Bush administrations, as well as Congress.

U.S. officials, however, yesterday denied that the meeting was intended to coordinate efforts to oust Syrian President Bashar Assad's government.

"That would be a monumental distortion," a senior State Department official said. "But it was a discussion about supporting reform and change in the region and specifically Syria – and how we can help that and work with people in the region and Syria to support that process."

The U.S. outreach is a direct result of President Bush's discussion last month with French President Jacques Chirac, said U.S and European officials. Advising against any discussion of "regime change," Chirac told Bush that the Damascus government was unlikely to survive the withdrawal of Syrian forces from Lebanon. The French president predicted that free elections in Lebanon would in turn force change inside Syria, possibly unraveling Assad's government, U.S. sources said.

Since that Feb. 21 meeting, the Bush administration has begun looking at possible political options in Syria, said analysts familiar with the U.S. thinking. "They're taking seriously that a consequence of getting out of Lebanon will be the collapse of the Assad regime, and they're looking around for alternatives," said Flynt Leverett, former senior director for Middle East affairs at the National Security Council under Bush.

The Syrian Americans who attended the meeting urged the administration to take tentative steps to pressure Damascus, such as having Bush call for greater freedoms and release of political prisoners, said Farid Ghadry, president of the Syrian Reform Party.

The delegation also sought support for lawsuits in U.S. courts against Syrian officials engaged in human rights abuses, an option available under the Alien Tort Claims Act, Ghadry said. The 1789 law grants jurisdiction to U.S. federal courts over "any civil action by an alien for a tort only, committed in violation of the law of nations or a treaty of the United States."

Ghadry said the Syrian opposition was encouraged by the "open and constructive" meeting, which was attended by key players in the administration's democracy policy such as John Hannah from Cheney's office, Robert Danin from the National Security Council and the Pentagon's David Schenker.

"They wanted to hear from us how they can help in extending the message of freedom and democracy in Syria," said Ghadry, who left his homeland 30 years ago, when he was 10, and formed his party after the Sept. 11, 2001, terrorist attacks. "They listened and

took a lot of notes. We felt from the responses that they understand these are important issues."

Some U.S. analysts and other Syrian Americans warned that the Syrian Reform Party and its allies are unrepresentative and too small to have any impact.

"Its membership is extremely thin and is not taken seriously. It's almost unheard-of in Syria," said Murhaf Jouejati, director of George Washington University's Middle East Studies Program.

On Lebanon, Rice said the United States is waiting to hear recommendations from U.N. envoy Terje Roed-Larsen on how to support spring elections there. "The main thing is just to help the Lebanese opposition and others, the entire Lebanese political space [and] people to get organized so that they can have a competitive, free and fair election," she said.

"I would suspect that if the U.N. comes back and says [do election] monitoring, people will be very supportive of that," Rice added. "Perhaps if there's need for nongovernmental organizations to do training or the kind of things that have been done in other places, I'm quite sure that people would be prepared to do that."

E.25 CNN March 27

A blast rocked Beirut on Saturday night, and Lebanese television broadcast stark images of severely damaged structures engulfed in flames.

Security officials reportedly think a car bomb caused the blast.

Casualties cannot yet be confirmed in the blast, which is said to have occurred in eastern Beirut, a predominantly Christian area.

Ambulances raced to the scene, reportedly in an industrial district, but it is not known whether people were trapped in the buildings.

A print shop and possibly a timber yard were said to be affected, which would account for the extensive flames and smoke.

News footage showed firefighters attempting to extinguish the fierce flames, which were shooting out of many windows.

The explosion was the latest in a series of blasts in the Christian areas of Lebanon. It occurred on the eve of the Christian holy day of Easter.

Four days ago, a bomb ripped through a shopping mall in a predominantly Christian area north of Beirut, killing three people and wounding at least two others, police said.

Just a few days earlier, a car bomb exploded in another Christian area of Beirut, shearing off part of a multistory office building. Nobody was killed.

The Lebanese capital had been relatively peaceful since the 1975-1990 civil war.

But the assassination last month of former Prime Minister Rafik Hariri has generated popular anger at Syria, which many think was behind his killing, and instability has re-emerged.

There have been large demonstrations against Syria's troop presence in Lebanon, and Prime Minister Omar Karami's government resigned.

Karami stepped down February 28 under intense pressure. But he was reappointed by parliament to bring together opposition and loyalist politicians in a Cabinet to lead Lebanon to general elections scheduled for May.

Syria began pulling its 14,000 troops to the Bekaa Valley near the border March 8, and vowed to bring all the troops and intelligence officials across the border into Syria.

A fact-finding team investigating Hariri's assassination has blamed Syria's government for the political tension that preceded the killing, according to a U.N. report released last week.

According to the report, the specific causes for the assassination will not be known until after the killers are brought to justice.

"However, it is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon and a failure of the Lebanese state to provide adequate protection for its citizens," investigators concluded.

The report called Lebanon's security services "negligent" and accused them of contributing to the "propagation of a culture of intimidation and impunity."

The 15-year civil war mostly pitted Lebanon's ruling conservative Christians against leftist Muslims, with Syria, Israel and Western international forces – including U.S. Marines – occasionally taking part.

The treaty that ended the fighting revised the constitution to give the Muslim majority a greater role. The presidency, chosen by parliament, goes to a Christian. The prime minister must be a Sunni Muslim.

E.26 CNN March 27

A bomb has exploded in a mainly Christian neighborhood of Beirut the night before Easter Sunday, setting off dramatic fires and wounding several people.

Local media said up to eight people were injured after what is believed to be a car bomb exploded in the industrial district Saturday night.

The attack is the third to hit Christian areas in Beirut since March 19 and is another blow to a nation with deep political divisions after the assassination of former Prime Minister Rafik Hariri.

Local media said up to eight people were wounded.

A print shop and possibly a timber yard filled with combustible materials were said to be in the area, accounting for some of the heavy smoke and fire.

News footage showed dramatic scenes of firefighters attempting, without much luck, to extinguish fierce flames shooting from windows.

The United States condemned the bombing in a written statement from State Department spokesman Adam Ereli.

"We call on the Lebanese authorities to exercise their responsibility to the Lebanese people to provide for their security and to identify and bring to justice those responsible for these acts," the statement said.

Last month's assassination of Hariri has provoked outrage over Syria's military and political influence, setting off massive demonstrations.

Many Lebanese believe Syria was behind the killing and Prime Minister Omar Karami and his government resigned on February 28 under intense pressure.

But he was reappointed by parliament to bring together opposition and loyalist politicians in a cabinet to lead Lebanon to general elections scheduled for May.

Syria began pulling its 14,000 troops to the Bekaa Valley near the border March 8, and vowed to bring all the troops and intelligence officials across the border into Syria.

A fact-finding team investigating Hariri's assassination has blamed Syria's government for the political tension that preceded the killing, a U.N. report released last week showed.

"It is clear that the assassination took place in a political and security context marked by an acute polarization around the Syrian influence in Lebanon and a failure of the Lebanese state to provide adequate protection for its citizens," investigators concluded.

A 15-year civil war mostly pitted Lebanon's ruling conservative Christians against leftist Muslims, with Syria, Israel and Western international forces – including U.S. Marines – occasionally taking part.

The treaty that ended the fighting revised the constitution to give the Muslim majority a greater role.

The presidency, chosen by parliament, goes to a Christian. The prime minister must be a Sunni Muslim.

E.27 New York Times March 27

A car bomb exploded in a predominantly Christian neighborhood in eastern Beirut on Saturday night, the eve of Easter, wounding at least five people and destroying two factories.

The bomb, which went off about 9:30 p.m. in the Sid al Bushriya neighborhood, was the third attack on a Christian neighborhood in a week.

Last Saturday, a bomb exploded in northern Beirut, wounding nine people. On Wednesday, two people were killed when a bomb exploded at a shopping center in Jounieh, north of Beirut.

The bombings have taken place against a backdrop of political turmoil set off by the assassination on Feb. 14 of Rafik Hariri, the former Lebanese prime minister.

Since then, hundreds of thousands of Lebanese have joined demonstrations blaming Syria and demanding that it withdraw its 10,000 troops from Lebanon.

Lebanon's Christians, who coexist with large populations of Sunni and Shiite Muslims and Druse, are overwhelmingly opposed to Syria's military occupation.

This week, a United Nations investigation concluded that heavy-handed Syrian interference in Lebanese affairs had contributed to the polarized political atmosphere that led to Mr. Hariri's assassination, and that a flawed local investigation had stymied efforts to find the killers.

The bomb exploded in a street ringed by three five-story buildings, destroying two of them and blasting gaping holes in the third. Fire engulfed two of the buildings, and the twisted wreckage of a car, presumably the carrier of the bomb, lay in the street.

The number of casualties was unclear. A Lebanese rescue worker at the scene said two people had been wounded or possibly even killed, while other security officials said five people had been wounded.

Given the damage caused, the bomb clearly could have caused a far higher number of casualties if the neighborhood, a dense urban area of factories and shops, had not been largely empty at the time.

Many of the Lebanese drawn to the site said they suspected Syrian involvement.

"The Syrians want to make a conflict with the Muslims," said Michel Kifrawi, a Lebanese Christian. "But they are not going to succeed. We want them to leave anyway."

E.28 New York Times March 27

In memory, the two scenes are linked by their silence. Last week in downtown Beirut, Lebanese by the hundreds filed past the tomb of Rafik Hariri, the fallen national leader, each pausing to offer some unspoken tribute. The only audible sound was a murmured prayer for the dead.

In Baghdad two months before, Iraqis in similar numbers had waited in line outside a high school to cast their ballots. Mortar shells were exploding in the distance, yet hardly anyone uttered a sound.

Amid such overwhelming displays of popular will, it seemed that words were hardly necessary.

Only weeks apart and a few hundred miles away, the popular demonstrations in Lebanon and Iraq offer themselves up for such comparisons. Their proximity suggests a connection, possibly one of cause and effect, like the revolutions that swept Eastern Europe in 1989. As went Berlin, Prague and Bucharest; so goes Baghdad, Beirut and Cairo.

President Bush has asserted as much, arguing that the toppling of Saddam Hussein and the holding of elections in Iraq set loose the democratic idea and sent the tyrannies reeling. From a distance, Lebanon looks like a domino.

Up close, though, it seems like something far more complex. For a correspondent who has spent much of the past two years inside Iraq, arriving in the seaside capital of Beirut is

a bracing and abrupt experience. For all the glories of election day, Iraq is still a grim and deadly place, where the traumas of the past 30 years are imprinted in the permanent frowns of ordinary Iraqis. Lebanon, by contrast, seems Iraq's sunny, breezy cousin, where young men arrive at demonstrations wearing blazers and hair gel, and the women high heels and navel rings. When the protest is finished, they drive off together in their BMW's.

How could Iraq have inspired this?

Chibli Mallat, a Beirut lawyer and opposition leader, has an answer. He believes that for years, Iraq stood as both a positive and malevolent symbol to others in the Middle East. Saddam Hussein's survival following the Persian Gulf war in 1991, Mr. Mallat said, froze the status quo in the region for more than a decade. The Iraqi dictator's prolific human rights abuses had the perverse effect of making every other unelected leader in the Middle East look tame by comparison. The result, he said, was political stasis.

"Saddam's survival created an atmosphere where people literally got away with murder," Mr. Mallat said. "His removal became a precondition for change in the region."

When the Americans finally returned to topple Mr. Hussein two years ago, and, more important, when millions of Iraqis risked their lives to cast ballots in January, the country emerged as a symbol for change across the region.

"Suddenly, there was a demand for democracy," Mr. Mallat said.

Mr. Mallat's view, compelling though it is, is a minority one in Lebanon. Most Lebanese will tell you that Iraq had nothing to do with the popular upheaval now gripping the country, and not just because they opposed the American invasion of their Arab neighbor. Unlike Iraq, Lebanon has been a functioning democracy since 1990, when the civil war, which killed 100,000 people, finally came to an end. Lebanon's press is vibrant, with newspapers and television stations largely free to criticize the government in Arabic, English and French. While Iraq still requires billions of dollars to repair its crumbling public works, Lebanon, thanks in no small way to Mr. Hariri's efforts, has largely rebuilt itself.

Indeed, it is no accident that the main slogan of the Lebanese opposition is not "Democracy," but "Sovereignty, Independence and Freedom." The goal is to expel Syrian forces, who have been in Lebanon for 30 years.

At least to an outsider, the main difference between Iraq and Lebanon seems not just Iraq's inexperience with democracy, but its all too dreadful experience with terror. In Iraq, political discourse often seems stunted, if less by a lack of practice than by the lingering shadow of Mr. Hussein. In Lebanon, with some exceptions - like the subject of Syria and its Lebanese client, President Emile Lahoud - most citizens are well accustomed to speaking their minds. In the last few weeks, most of the remaining taboos have fallen away.

"We want the truth," said Naila Shukry, a biology student at Arab University in Beirut. "Someone has murdered our leader, and we want to know who is responsible."