Understanding the process of multi-document summarization: content selection, rewriting and evaluation

Ani Nenkova

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

COLUMBIA UNIVERSITY

2006

ABSTRACT

Understanding the process of multi-document summarization: content selection, rewriting and evaluation

Ani Nenkova

Recent years have seen unprecedented interest in news aggregation and browsing, with dedicated corporate and research websites becoming increasingly popular. Generic multi-document summarization can enhance users' experiences with such sites, and thus the development and evaluation of automatic summarization systems has become not only research, but a very practical challange. In this thesis, we describe a general modular automatic summarizer that achieves state of the art performance, present our experiments with rewrite of generic noun phrases and of references to people, and demonstrate how distinctions such as familiarity and salience of entities mentioned in the input can be automatically determined. We also propose an intrinsic evaluation method for summarization that incorporates the use of multiple models and allows a better study of human agreement in content selection. Our investigations and experiments have helped us to understand better the process of summarization and to formulate tasks that we believe will lead to future improvements in automatic summarization.

It is well-known that humans do not fully agree on what content should be included in a summary. Traditionally, this phenomenon has been studied on the level of sentences, but sentences are a rather coarse level of granularity for content analysis. Here, we introduce an annotation method for semantically driven comparison of several texts for similarities and differences on the subsentential level. When applied to human summaries for the same input, the method allows for a better examination of human agreement, and also provides the basis for an evaluation method that incorporates the notion of importance of a content unit in a summary.

Given the variability of human choices, we next address the questions of what features in the input are predictive for inclusion of content in the summary. We use a large collection of human written summaries and the respective inputs to study the predictive effect of one feature that has been widely used in summarization: frequency of occurance. We show that content units that are repeated frequently in the input tend to be included in at least some human summaries and that human summarizers tend to agree more on the inclusion of frequent content units. In addition, human summaries tend to have higher likelihood under a multinomial model estimated from the input than automatic summaries do. This empirical investigation leads us to propose an algorithm for a context sensitive frequency-based summarizer. We show that context sensitivity and a good choice of composition function for estimating the weight of a sentence lead to a summarizer that performs as well as the best supervised automatic summarizer.

We then turn to exploring methods for summary rewrite; that is, techniques for automatic modification of the original author's wording of sentences that are included in a summary. The added flexibility of subsentential changes has potential benefits for improving content selection as well as summary readability. We show that human readers prefer summaries in which references to people have been rewritten to restore the fluency of the text. We further develop our work on references to people, by presenting an approach to automatic classification of entity salience and familiarity, based on robustly derivable lexical, syntactic and frequency features. Such information is necessary for the generation of appropriate referring expressions.

Contents

List of Figures

List of Tables

Acknowledgments

It is an exciting time for me to be finishing my dissertation and life as a student. I owe thanks to many people who turned my years at Columbia and New York City into a wonderful experience.

Above all, I would like to thank my amazing adviser, Kathy McKeown. She has been and will be a role model for me, and taught me a lot about focus, determination and life as a researcher. I am grateful to her for always supporting and encouraging me, while at the same time giving me the freedom to explore new directions and topics. I would also like to thank Julia Hirschberg for being an irreplaceable mentor and friend during the past three years. It was so good to know I can always stop by her office and chat, and share my problems, joys and ideas. I would like to also acknowledge the rest of my committee members—Steve Feiner, Kathy McCoy and Becky Passonneau, for their helpful comments and suggestions.

I was lucky to be able to work with many people during my years at Columbia, and I have learned a lot from my collaborators. I am especially grateful to Becky Passonneau and Lucy Vanderwende for the interesting discussions we had, and for their support and encouragement. It was also great to work with Advaith Siddharthan, he was a wonderful friend, and with his humor and great spirit turned conference deadlines into a fun event. I will not hold against him his consistent domination in table tennis and tennis. I also owe many thanks to Barry Schiffman, with whom I co-authored my first paper at Columbia. I will miss our tea breaks.

It was great to be part of the NLP group at Columbia, and I thank all the members for their help, interesting discussions, presentation feedback and good time. I was fortunate to be part of such a diverse and numerous research group, as well as to be surrounded by wonderful officemates, students, professors and administrators from the computer science department. And finally, many thanks to my dear friends in New York and Sofia, who kept me sane and made me laugh.

To my grandfather Paskal

Chapter 1

Introduction

Recent years have seen unprecedented interest in news aggregation and browsing, with dedicated corporate and research websites becoming increasingly popular: news.google.com, newsbot.msn.com, news.com, newsinessence.com, newsblaster.cs.columbia.edu. Generic multi-document summarization can enhance the users' experiences with such sites, and thus the development and evaluation of automatic summarization systems has become not only research, but a very practical challenge.

Summarization is a challenging task for automation because, when different humans summarize the same articles, they include different content from each other, reflecting their personal interest and background knowledge. This fact also poses a challenge for summarization evaluation, which is traditionally done through a comparison between the system output and a model produced by a human. In this thesis, we take an empirical approach to tackling these challenges. We use the summarization corpora that have been made available in recent years to quantify human agreement and to use human agreement to assign differential weights to different information content. The idea we use is the following: the more humans agree on the inclusion of certain content, the more evidence we have that this content is objectively important for inclusion in a summary. We incorporate this in a diagnostic and reliable evaluation metric based on multiple models. Furthermore, our analysis of human performance allows us to determine the features that predict good performance.

Such features should be predictive of human agreement, characterizing content that several humans will include in their summary. We use this approach to identify useful summarization features, and to build and evaluate a modular summarizer. Such an empirically grounded approach to system development has not been done in the past, partly due to the previous lack of a large enough corpus with the necessary data consisting of multiple human models and summarization inputs.

Another characteristic of human summaries is the high degree of reformulation of the text from the input to form a fluent abstract. Human summarizers do not simply pick sentences from the input: some sentences in the input may contain trivial or repetitive information alongside very important information. So, in order to include the best content in their summary, people choose important pieces and combine them together in new sentences. The process of abstraction also avoids readability problems that can arise if summarizers are constrained to picking entire sentences from the input. The form and level of details of references in the sentences depend on the context they appear in, and might not be appropriate for the new context created in the summary. Obvious examples of such problems are the inclusion of pronouns or definite noun phrases, the full interpretation of which depends on the preceding text. It is thus important for automatic summarizers to make use of techniques that rewrite the original text from the input in order to improve the content selection and the readability of the final summaries. In this thesis, we develop summary rewrite techniques for noun phrases and for references to people. In both types of rewrite, context plays an important role, determining the most appropriate form of specific references.

A final observation about summarization of news that we exploit in this thesis is that news often revolves around people and both the inputs to the summarizer and the summaries contain many references to people. This is important, because it allows us to develop more specific models for references to people that are applicable to domain-independent news summarization. In this thesis, we develop a model for the appropriate flow of syntactic realizations of subsequent mentions to people. We also develop robust classifiers to determine the familiarity and importance of people mentioned in the the input, and demonstrate how these distinctions help to rewrite references to forms similar to those that a human summarizer would use.

We now give a brief guide to the content presented in each chapter of this thesis and then outline our main contributions.

1.1 Thesis organization

In chapter 2, we begin our exploration of multi-document summarization by presenting the motivational results of a user study, which shows that summarization can help information-seeking users in a news browsing site, both by making their experience more pleasant and by helping them find more relevant information. We then proceed to a comparison between multi-document summarization and single-document summarization, which has a much longer tradition. We see that there is more agreement on content selection between human summarizers in single- than in multi-document summarization. We then proceed to characterize the differences in content selection in terms of difference of vocabulary. Since it is possible that differences in vocabulary usage across different summarizers might be due to the use of synonyms and paraphrases, we compare the differences with those observed in multiple translations of the same text. We demonstrate that the variability in sets of human summaries for the same input far exceeds the variability that can be due to alternative expressions of the same content. We thus postulate that different probabilities for emission into a summary are associated with different content in the input.

In Chapter 3, we develop a content unit annotation procedure that allows us to compare different human summaries and confirm the probabilistic emission hypothesis. We observe that content units have a Zipfian distribution (Zip65), a distribution characteristic of complex optimization of multiple constraints problems. We use the annotation to estimate which content units have higher emission probabilities, or weights, than others; these are the content units that appear in more human summaries. We then propose an evaluation method that incorporates these weights, leading to stable evaluation results that do not depend on the choice of model and that predict that multiple equally good summaries for the same input are possible. We discuss the application of the evaluation method in a large-scale evaluation, including annotation reliability between novice annotators and correlations with other automatic and manual evaluation metrics.

In Chapter 4, we present a context-sensitive frequency-based summarizer and show how its modularity allows us to integrate noun phrase rewrite, which permits more flexibility in content selection than an approach based on sentence extraction. The summarization algorithm we developed was motivated by the analysis of multiple human summaries.

Our context-sensitive summarizer is designed so that specific characteristics of summarization can be studied. The first aspect of the summarizer that we evaluate is based on the idea that the importance of content is compositional. Basic units of meaning, such as content words, can be assigned importance, and then an appropriate composition function is necessary to combine the importance of these basic units into importance weights for larger units, such as sentences. For our summarizer, we use frequency in the input to assign importance weights to content words. We demonstrate how frequency in the input is related to human content selection, with humans being likely to include frequently repeated content in their summaries, and likely to agree on inclusion of frequently repeated content. But more importantly, we demonstrate that the choice of composition function for assigning weights to sentences has an enormous impact on the performance of the summarizer. Some choices lead to performance that is only as good as the baseline, while others lead to performance as good as that of the current best summarizer tested on a large scale common test set evaluation.

Another major aspect of a summarizer that we evaluate is its sensitivity to context. We study how context (i.e., previous selections) can impact subsequent content selection choices. At a minimum, a summarizer needs to avoid repetition in the summary, which potentially could easily occur, given that the input to the summarizer is characterized by a high degree of redundancy. Our experiments confirm this intuition, and show that a sensitivity to context leads to overall better content selection and significant reduction of repetition in the summary.

Our method clearly shows the steps that lead to good summarizer performance, and the summarizer that we develop based on these observations performs as well as the state-of-the-art summarizer from recent large-scale evaluations both in content selection and readability. Moreover, our method is unsupervised, does not require any background corpora and does not involve empirically set parameters.

The combination of context sensitivity and compositional assignment of weights does not constrain us to sentence extraction. We use the same approach for noun phrase rewrite, allowing more flexibility in content section. The importance of alternative possible noun phrase realizations, containing different amount of detail, is evaluated depending on the context of previous selections. In this entity-centered approach, the best realization is chosen for each main entity in a sentence. The approach leads to about 50% change in the summary content compared to the purely extractive version of the same summarizer. These results show that context sensitivity is very important not only for extractive content selection, but also for rewrite. Moreover, our entity-centered approach leads to summaries that are significantly better in content selection, grammaticality and referential clarity than an event-centric generation summarizer that was tested on the same data.

To obtain more semantically and linguistically motivated information from the summaries without compromising the domain independence of the summarizer, we turn to the study of references to people. By narrowing rewrite to reference people, we can obtain greater benefits because the chance of grammatical errors is reduced, while still improving readability and content selection. Our results are reported in Chapter 5. There, we show how narrowing the focus down to references to people, we can develop models of the flow of subsequent references, as well as build automatic classifiers for recognizing the familiarity and salience of people referenced in the input to a summarizer. We also demonstrate how such distinctions can be used for rewrite of first and subsequent mentions to people, which readers find more natural and which reproduce human reference generation decisions with high accuracy.

1.2 Thesis contributions

In this thesis, we make contributions to all three aspects of summarization: content selection, summary rewrite for changing the original input text on a subsentential level, and empirically grounded evaluation. The extensive analysis of human summaries that we perform allows us to study the characteristics of human summarization behavior, and to define an evaluation method for summarization that captures these characteristics. In addition, the analysis of human summaries motivates the features and system development of our compositional context sensitive summarizer. We demonstrate that context sensitivity, which is not traditionally a focus in summarization research, is critical both for content selection and noun phrase rewrite, and also plays a role in determining the appropriateness of references to people in summaries.

Content selection: composition function and context We have developed a system that exploits the advantages given by context sensitivity and compositional derivation of importance. The clearly defined modularity of the system allows us to assess the contribution to performance of each component, in contrast to most other summarizer architectures proposed in the past. Our unsupervised system achieves performance equal to that of the state-of-the-art supervised approach.

Noun phrase rewrite We have developed and evaluated an approach to entity-centered rewrite of generic noun phrases. The approach uses the techniques we successfully developed for extractive summarization, to combine grammaticality and importance in a given context to choose the best among several noun phrase alternatives. The approach leads to summaries that are 50% different from those produced by a purely extractive summarizer, and significantly outperforms a fully generative event-centered summarizer in content selection capabilities, as well as in grammaticality and referential clarity.

Rewrite of references to people References to people are ubiquitous in news, leading to the necessity of a good theory for references to people. In this thesis, we have demonstrated that distinctions such as salience and familiarity of the referent can be robustly determined from features in the input and that they determine the appropriate form of reference. Our results show that readers prefer rewritten summaries, and that we can reproduce human decisions for generation of first references with high accuracy.

Evaluation using multiple models The variation in human content selection behavior has led to the need to develop novel evaluation methods, especially necessary for abstractive summarization. We have developed an annotation procedure for semantic, subsentential, comparison of a set of summaries on the same topic, allowing us to assign deferential weight to different content and permitting a better study of human choice variation compared to previous restrictive sentence-based approaches. We proposed an evaluation approach that uses multiple models and the empirical weighting of information, thus avoiding model bias.

Chapter 2

The process of multi-document summarization: analysis of human and automatic behavior

Interest in multi-document summarization of newswire started with the advent of on-line publishing and the increased impact of the Internet. Research on the topic builds upon a long tradition established in the area of single-document summarization (Luh58; RRS61), but the first automatic multi-document summarizers were developed only in the mid 90s (MR95; GMCK00; ABN00) and there has been growing interest in the field ever since. In the past decade, many automatic summarizers for news have been developed and multi-document summarization has often been the main topic in dedicated summarization workshops. In particular, the Document Understanding Conference (DUC) was established in 2001 as a forum for evaluating system performance on common datasets and multi-document summarization has been one of the tasks in all years from 2001 to 2005.[1] These conferences have produced a wealth of data for multi-document summarization research—input sets of multiple articles for summarization, human abstracts and extracts for these inputs, as well as summaries produced by the participating automatic systems. I will use these data throughout the thesis for analyses of trends, and for comparison and evaluation.

[1] Online proceedings of the annual conferences are available at `http://duc.nist.gov`

In this chapter, I will motivate the work in the rest of the thesis by overviewing advances in multi-document summarization, discussing a task-based evaluation in which multi-document summaries are found to enhance users' experience during a report writing task (section 2.1). I will also compare multi- and single-document summarization results in past DUCs, showing that the single-document task is easier, with humans agreeing more on content selection decisions; at the same time, more progress has been made in multi-document summarization, where automatic systems outperform the baseline, in contrast to single-document summarizers (section 2.2). Finally, I will analyze the data collected in DUC to understand better the process of summarization, as well as to identify problematic areas for automatic summarizers. The main findings of the analysis will be used to formulate specific hypotheses and tasks that will be studied in-depth in the later chapters of this thesis. To preview some of the findings:

1. The analysis of automatic summaries shows the need for improving their readability. The analyses presented here (section 2.3) indicate that the majority of automatic summaries have very poor to barely adequate referential clarity, focus and structure, and coherence. This points to the need to evaluate the readability aspects of a summarizer, not just its content selection capabilities as has been more traditionally done, and to develop systems that incorporate readability considerations. In chapter 5, I will discuss *an approach for summary rewrite* of references to people that can lead to improvements in readability, and an approach to rewrite of noun phrases will be presented in chapter 4.

2. One of the characteristics of summarization in general and multi-document summarization in particular, is that humans do not choose the same content for their summaries of the same input (section 2.4). There is thus a need to establish a protocol that will allow the study of this phenomenon in units of the right granularity, larger than content words and smaller than sentences, as well as to incorporate this observed variability into an evaluation measure for abstractive summarization. In chapter 3, I will outline an *annotation procedure* that facilitates the study of similarities and differences between multiple summaries, and which is used as the *basis for an evaluation*

protocol for content selection.

3. The fact that different people choose somewhat different content for summaries impacts the formalization of the summarization task. It, for example, suggests that it will be difficult to reliably cast summarization as a simple binary classification problem in which for each sentence in the input the summarizer decides if it should be selected for the summary or not. Different humans will annotate the training data differently, and it is not clear how this would impact the performance of the summarizer. Our analysis here (section 2.4) suggests that a probabilistic model corresponds better to empirical observations about content selection, leading to a proposal for a novel *formalization of the process of summarization* for generic multi-document summaries, to be fully described in chapter 4. There we will discuss the applicability of a multinomial model over content words that is estimated from the input and an automatic summarizer based on the model.

We now turn to our overview of progress in summarization and to the detailed analysis of content in human-authored summaries, content in machine summaries and the assessment of readability of machine summaries, all based on data from DUC.

2.1 Do generic summaries help?

The overall topic of this dissertation is generic multi-document summarization. It is thus appropriate to begin by motivating our work and showing how real users can benefit from such summaries. Here I briefly discuss a task-based evaluation showing that generic summaries can help users writing a report to produce better reports and to make their experience with a news browsing system more pleasant.

The automatic summaries used for the evaluation were produced by Newsblaster—a system that provides an interface to browse the news, featuring multi-document summaries of clusters of articles on the same event (MBE$^+$02; MBC$^+$03). The users were also exposed to alternative interfaces: one featuring no summaries at all, and one featuring a first sentence summary for each document, as well as a one-sentence centroid summary for each cluster. Human subjects used the different interfaces (with no access to other information but that provided through the interface) to write reports answering three questions for each of three events. For each event, there were four clusters of about 10 news articles—two directly related to the topic and two peripherally related. The human subjects were asked to collect facts that answer the questions they were given, and after completing each report they were also asked about their experience writing the report. There were 13 subjects writing reports using an interface with no summaries, 11 using lead-sentence summaries, and 10 using Newsblaster summaries.

The reports written using Newsblaster summaries were *significantly* better than those written using an interface with no summaries. In addition, the subjects felt they had enough time to write the report, they read less of the source documents and felt that the reporting writing task was easier when they had Newsblaster summaries than when no summaries were provided.

The full details of the study are described in (MPE$^+$05). The results of the study confirm that generic summarization is a useful task and that automatically produced summaries can lead to improvements of report quality and user satisfaction, motivating the need for summarizer development. In chapter 4, I will discuss a generic summarizer for news based on probability distribution assumptions about content.

2.2 Comparison between multi- and single-document summarization

Task-based evaluations such as the one described in the previous section are difficult to plan and execute and because of this, intrinsic evaluation is usually performed, where the quality of summaries is measured directly. In the Document Understanding Conferences, an intrinsic evaluation of summaries was performed in terms of *coverage*, that is the degree of overlap between a summary and a human written model summary. We now use this metric to overview the progress in the area and compare the characteristics of single- and multi-document summarization.

2.2.1 DUC coverage metric

Before describing the results of our comparison, we describe how the *coverage* metric was defined and used in DUC.

1. A human subject reads the entire input set and creates a 100 word summary for it, called a model.

2. The model summary is split into content units, roughly equal to clauses or elementary discourse units (EDUs). This step is performed automatically using a tool for EDU annotation developed at ISI.[2]

3. The summary to be evaluated (a peer) is automatically split into sentences. (Thus the content units are of different granularity—EDUs for the model, and sentences for the peer).

4. Then a human judge evaluates the peer against the model using the following instructions: For each model content unit:

   (a) Find all peer units that express at least some facts from the model unit and mark them.

   (b) After all such peer units are marked, think about the whole set of marked peer units and answer the question:

   (c) "The marked peer units, taken together, express about k% of the meaning expressed by the current model unit", where k can be equal to 0, 20, 40, 60, 80 and 100.

The final summary coverage score is based on the content unit coverage. In the official DUC results tables, the score for the entire summary is the average of the scores of all the content model units, thus a number between 0 and 1.

[2] http://www.isi.edu/licensed-sw/spade/.

2.2.2 Comparison of summarization tasks

In order to study the differences between summarizers, we fit a two-way analysis of variance model with the systems and the input as factors and the coverage score as dependent variable. The simple main effect model we use is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{2.1}$$

Y_{ij} is the coverage score for system i for input j, and it is equal to μ, a grand mean coverage for any summary, adjusted for the effect of the summarizer (α_i) and the effect of the input (β_j) and some random noise ϵ_{ij}. Other possible sources of variation such as evaluator, or system/evaluator interactions and the like, are not included in the model, since they have been controlled for in the experimental design as discussed in (HO04).

For all tasks, both main effects are significant with $p = 0$, which indicates that significant differences between summarizers exists, as well as that some sets or documents are easier to summarize than others (and summarizers get higher coverage scores for them). It is of interest to be able to compare each two summarizers against each other, total of $\binom{N}{2}$ comparisons when N is the number of summarizers. Thus, the number of pairwise comparisons to be made is quite large, since each year more than 10 automatic summarizers were tested and several human summarizers wrote summaries for comparison purposes. In order to control the probability of declaring a difference between two systems to be significant when it is actually not (Type I error), we use a simulation-based multiple comparisons procedure that controls for the overall experiment-wise probability of error. 95% confidence intervals for the differences between system means are computed with simulation size = 12616. Pairs of summarizers where the confidence interval for the difference between the two means excludes zero can be declared to be significantly different.

Tables 2.1 and 2.2 show the average coverage performance for single and multi-document summarizers for DUC 2002 (more detailed tables and discussion can be found in (Nen05)). Letter codes are assigned to human summarizers, and number codes to automatic summarizers. System 1 is the baseline for single document summarization—the first 100 words of the input article. System 2 is the baseline for multi-document summarization—the first n words of the latest published article (for $n = 50, 100, 200$).

summarizer	1	15	16	17	18	19	21	23	25	27	28	29
coverage	*.37*	.33	.30	.08	.32	*.39*	.37	.34	.29	.38	.38	.36
sums	291	295	295	292	294	295	295	289	294	292	295	294

summarizer	30	31	A	B	C	D	E	F	G	H	I	J
coverage	.06	.36	.47	.51	.46	.52	.49	.47	.54	.57	.53	.47
sums	294	292	30	30	24	30	30	25	25	30	29	30

Table 2.1: Summarizer code, coverage and number of summaries for the single document summarization task at DUC 2002. The last row gives the number of summaries produced by each summarizer.

summarizer	16	19	2	20	24	25	26	28	29	3
50 wrds	.12	.23	.14	.12	.13	.10	.21	.21	.16	.14
100 wrds	.13	.24	.13	.16	.19	.12	.19	.23	.14	.20
200 wrds	.15	.25	.13	.21	.22	.16	.24	.23	.17	.22
sums	59	59	59	59	59	59	59	59	59	59

summarizer	A	B	C	D	E	F	G	H	I	J
50 wrds	.37	.44	.25	.40	.46	.33	.44	.32	.26	.22
100 wrds	.37	.43	.38	.28	.25	.28	.36	.31	.39	.26
200 wrds	.37	.42	.45	.39	.27	.30	.29	.37	.41	.26
sums	6	6	5	6	6	5	5	6	6	6

Table 2.2: Coverage for summarizers at the multi-document task at DUC 2002, for different target summary lengths. The last row shows the number of summaries evaluated for each summarizer.

Multiple comparisons based on the simulation method allow us to draw the following conclusions:

1. Single document summarization systems do not significantly outperform the baseline, while multi-dococument summarization systems do. We can see from tables 2.1 and 2.2 the performance of the baseline and the best system. For single document summarization, the baseline (peer 1) performs at 0.37 and the best automatic summarizer at 0.39 (peer 19). For multi-document summaries of the same length, the baseline (peer 2) achieves performance of 0.13 and the best system (peer 28) outperforms it at 0.23, showing a five times bigger difference than in single-document summarization.

2. Differences between multi-document summarization systems become significant when longer summaries are produced. Overall, the performance of systems tends to improve as the summary size gets larger (see the columns in table 2.2), while the baseline performance remains the same for different lengths. But as summaries get longer, and with better content, more readability problems can be expected.

3. There is more human agreement on content in single document summarization than in multi-document summarization. We can see from the human performance figures in the two tables that multi-document summarization is harder than single document summarization in terms of predicted human agreement, with average human coverage of 0.503 for single-doc and 0.331 for multi-document human summarizers, both for 100 word summaries.

In summary, the discussion in this chapter has shown more evidence that multi-document summarization is a promising research area that can lead to competitive automatic systems that outperform simple baselines. The observed gap between human and automatic performance also indicates that there is much room for future research and improvement.

2.3 Readability analysis of machine summaries

In the previous section we discussed the coverage performance of automatic systems. The longer the target summary size, the better the system's performance (this trend is not manifested so strongly with human summaries). Also, there are more significant differences between systems at longer target summary lengths. But as the summaries get longer, more readability concerns will be raised.

For this reason we now take a look at the linguistic quality judgments for the summaries submitted to DUC 2005 (250 words). We developed five questions pertaining to linguistic quality and NIST assessors used them to judge the 1,600 automatic and 311 human summaries from the DUC 2005 evaluation. The degree to which each summary possessed a given quality was assessed on a scale from 1 to 5 (corresponding to "very poor", "poor", "barely adequate", "good", "very good").

The following quality description were used:

(Q1) Grammaticality The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments and missing components) that make the text difficult to read.

(Q2) Non-redundancy There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g.,"Bill Clinton") when a pronoun ("he") would suffice.

(Q3) Referential clarity It should be easy to identify to whom or what the pronouns and noun phrases in the summary are referring. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

(Q4) Focus The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

(Q5) Structure and Coherence The summary should be well-structured and organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Tables 2.3 and 2.4 give the fraction of summaries falling in each quality category for automatic and human summaries respectively. Even a cursory look at the two tables shows

	very poor	poor	barely adequate	good	very good
Q1	0.0775	0.0737	0.1512	0.3875	0.3100
Q2	0.0281	0.0343	0.0931	0.1918	0.6525
Q3	0.1462	0.2187	0.2743	0.2018	0.1587
Q4	0.1025	0.1968	0.3193	0.2387	0.1425
Q5	0.3006	0.4012	0.1768	0.0812	0.0400

Table 2.3: Fraction of automatic summaries into each quality category for the five linguistic properties. A total of 1,600 250-word summaries were evaluated

	very poor	poor	barely adequate	good	very good
Q1	0.0032	0.0000	0.0160	0.1446	0.8360
Q2	0.0000	0.0000	0.0192	0.0578	0.9228
Q3	0.0000	0.0000	0.0096	0.0450	0.9453
Q4	0.0000	0.0032	0.0225	0.0546	0.9196
Q5	0.0000	0.0096	0.0321	0.1350	0.8231

Table 2.4: Fraction of human summaries in each quality category for the five linguistic properties. A total of 311 250-word summaries were evaluated

that automatic summaries are definitely inferior to human summaries in terms of linguistic quality. More than 95% of human summaries fall in the “good” or “very good” categories for each question. In contrast, most of the machine summaries fall in the “very poor”, “poor” and “barely adequate” categories for the questions concerning referential clarity (Q3), focus (Q4) and coherence (Q5). These categories cover 64%, 62% and 88% of the automatic summaries respectively for Q3, Q4 and Q5.

The analysis shows that linguistic quality is an aspect of summarization in which a lot of improvement is possible. In section 5, I will outline an approach for summary rewriting to start addressing the problem.

2.4 Content in human-authored summaries

In addition to producing a readable fluent text, a summarizer needs to actually choose important and representative information to include in the summary. How do humans choose what content to include in a summary based on their reading of a document(s)? It is obvious that the answer to this question would have a big impact on the approach to developing automatic summarizers. The earliest research in summarization (Luh58) studied extractive summarization—the process of forming a summary by choosing sentences from the input, without reformulation. In this context it made sense to reduce the question of content selection to a binary decision—"Will a given sentence be included in the summary or not?" Such a view of the process of summarization is temptingly simple: a gold-standard can be created by a human and subsequently systems can be developed or trained to reproduce the gold-standard human selections. But the Rath *et al.* study (RRS61) revealed that different humans sometimes select different sentences for their summaries, and the same person can make different choices if asked to do the task twice, with about a week between the two selections.

The "humans don't agree" refrain will be reiterated over the years by many other researchers. But a closer inspection of the phenomenon shows that rather than leading to despair, this fact of summarization behavior simply leads to a picture of human summarization mechanism different from a sequence of binary deterministic decisions.

We hypothesize that a more plausible explanation of the mechanism of summarization would be a probabilistic (non deterministic) process according to which the content from the input, say sentences, has some probability of being emitted in a summary. Informally, one can imagine that after reading the documents to be summarized, people implicitly identfy units of content— content words, sentences, or some other granularity. Each of them is associated with an *emission probability*, P_n, which indicates how likely it is for a human to choose the unit of content in a summary. A few of the units of content will have high probability in comparison to others, some will never appear in a summary, and a large part of the units of content will have a relatively low emission probability. We will explore the applicability of this hypothesis for building a full-fledged summarizer in chapter 4, but in the remainder of this section we will present the empirical evidence for the validity of the proposed model.

Given multiple human summaries, we could estimate the probability distribution for the emission probabilities of units of content. The simplest carriers of content in a text are *content words* and for this initial investigation we will be constrained to studying the distribution of content words.

Definition Let w be a content word appearing in a human summary. A *summary frequency class*, C_n, consists of all content words that appear in exactly n of the summaries: $w \in C_n \longleftrightarrow w \texttt{ appears in } n \texttt{ human summaries}$.

We compute the size of the frequency classes in 7 human summaries of around 250 words each for 20 input sets (see figure 2.4). As expected, C_1, the summary frequency class containing words that are unique to only one of the seven summaries, is the largest frequency class for all 20 sets, and the size of the different classes is quite stable across the 20 sets. For a specific example, one can look at set$_1$ as a typical case, the sizes of frequency classes for which are shown in the first line of table 2.4. 604 content words appear in the 7 different summaries for this set (this is the sum of the first row in the table), and of these 381, or 63%, appear in one summary only (the last column in the table), compared to 20 (3%) content words that appear in all 7 summaries. We might wonder if we observe so little overlap because of the use of paraphrase and synonyms. We will show that the observed disagreement is bigger than what could be explained by alternative phrasing. Also, in the next chapter, we will look at the agreement between the humans in terms of manually marked content units, and there we will observe the exactly the same distribution.

So, the question arises about how much of these observations are specific to summarization, and how much can be explained with general lexical distribution properties. It is known that in general in a large corpus many lexical items will appear only once, i.e, in one document. In order to see if the observed distribution from summaries describes a peculiarity of summaries rather than a general phenomenon, the same statistics were computed for sets of 7 translations for 20 documents, each of roughly 250 words (shown in table 2.4). The size of frequency classes from the translation corpora can give us a reliable baseline for the expected differences in distribution that are due to general lexical frequency properties and to vocabulary variation accounting for paraphrase. This is because all seven translations express the same meaning, so the difference in vocabulary between the human translations describes the differences that are due to the range of possibilities for lexical realization. A significantly higher variation in summaries than in translations can be taken as a sign of differences in summary content, indicating that different people indeed express different content in their summaries.

The probability of a word appearing in C_1 (remember that C_i is the class of words used by i summarizers/translators), computed over the 20 respective sets, is 0.3334 for the summaries and 0.1730 for the translations. Analogously, the probability of a word appearing in class C_2 is 0.0942 for the summaries and 0.0521 for the translations. For both classes, the probabilities for summaries are significantly higher than those for translations according to a Welch Two Sample test on the vectors of 20 probabilities for the different sets ($t = 13.3339$, p-val < 0.0001; $t = 12.9419$, p-val < 0.0001 respectively). We focused on the classes C_1 and C_2 because they are the classes indicating highest disagreement between humans, C_1 specifically contains the words that were used by only one human. Both of these classes are larger and words fall in them more often in summarization than in translation. If we focus on the other end of the agreement spectrum, we see the opposite: the probability of a word in C_7 is significantly higher in translation than in summarization (0.0769 vs. 0.0567).

An additional experiment to confirm that summarization is characterized by large number of rare events in terms of lexical items is to see how much the vocabulary increases with the addition of new summaries. On average across the 20 sets, a single summary has 153 content words, and the collection of all 7 summaries of the same texts has an average vocabulary size of 580. For comparison, a single translation had on average 132 content words and the collection of 7 translations had 339 words—the vocabulary growth in summaries is significantly larger than can be explained by lexical variation and must be due to differences in content matter. Figure 2.1 shows graphs of the vocabulary growth for the 20 collections of 7 summaries/translations. The lines for summaries are drawn in solid red and those for translations, in dashed blue. All of the lines corresponding to summary collections are steeper than any given line corresponding to the rate of vocabulary growth in the set of seven translations, confirming that different human summaries indeed have variations that

SET	$\|C_7\|$	$\|C_6\|$	$\|C_5\|$	$\|C_4\|$	$\|C_3\|$	$\|C_2\|$	$\|C_1\|$
set_1	20	16	14	25	51	97	381
set_2	15	9	17	31	49	108	330
set_3	13	14	15	29	42	90	363
set_4	16	9	15	24	54	104	347
set_5	25	19	31	29	36	67	254
set_6	12	8	12	19	49	110	441
set_7	9	15	27	29	59	123	360
set_8	7	9	15	31	57	105	480
set_9	12	10	16	32	46	101	440
set_{10}	14	16	24	35	46	106	304
set_{11}	11	10	11	30	51	106	357
set_{12}	17	13	17	30	44	97	367
set_{13}	10	12	19	31	49	100	463
set_{14}	18	15	16	24	42	104	308
set_{15}	17	13	13	23	55	107	336
set_{16}	18	7	16	29	44	94	294
set_{17}	14	11	12	24	48	96	339
set_{18}	12	20	16	33	60	84	266
set_{19}	14	13	13	25	39	106	417
set_{20}	13	13	9	27	41	125	359

Table 2.5: Each entry shows the size of the summary frequency class for 20 input sets. C_i is the class of words that appear in i summaries. There are 7 human summaries for each set and thus C_7 is the highest class. As expected under our hypothesis, words predominantly appear only in one or two summaries.

SET	$\lvert C_7\rvert$	$\lvert C_6\rvert$	$\lvert C_5\rvert$	$\lvert C_4\rvert$	$\lvert C_3\rvert$	$\lvert C_2\rvert$	$\lvert C_1\rvert$
set_1	39	25	27	26	24	43	183
set_2	39	26	29	29	36	57	178
set_3	35	20	20	25	21	53	131
set_4	31	35	14	28	24	46	187
set_5	48	16	15	19	21	48	125
set_6	34	25	22	15	21	28	106
set_7	40	25	18	24	44	45	155
set_8	32	26	20	21	28	61	174
set_9	27	17	25	33	31	57	214
set_{10}	45	24	13	18	21	30	123
set_{11}	30	28	28	25	33	46	185
set_{12}	41	26	21	23	38	48	211
set_{13}	36	18	12	22	24	59	127
set_{14}	37	29	10	16	21	38	137
set_{15}	25	26	11	28	42	52	191
set_{16}	40	15	16	23	30	43	169
set_{17}	40	31	26	19	29	59	165
set_{18}	31	23	18	22	24	49	173
set_{19}	35	15	17	10	27	48	134
set_{20}	44	20	16	26	37	52	145

Table 2.6: Class frequency for 7 human translations of 20 texts. The distributions are quite stable across the 20 sets and there are significantly fewer words classes C_1 and C_2 (that is words that appear in only one or two translations) in comparison to summarization.

are due to differences in content, rather than just to lexical choice variation.

But content words might seem too small a unit for analysis. Another possible granularity is that of a sentence—we could ask if the hypothesized model of summarization from the beginning of section 2.4 holds at the sentence level as well. In (RT03), the question of the degree to which a given sentence is likely to be included in the summary was studied in the context of developing an approach to summarization evaluation. An assessor was asked to rank *each sentence from the input* on a scale from 10 to 0, where a mark 10 means that the sentence is highly relevant and suitable for inclusion in a summary, while a mark of 0 means that the sentence should never be included in a summary. This way of assigning importance is different from the one we used for words, based on the observed use by a human summarizer. Unfortunately, no data for sentence selection summaries by multiple humans are available, even though this would be the best data for our study. We use direct rankings here as an approximation and specifically focus on sentences with ranking 5 or more, since accoring to the instructions given to the people assigning the score, any sentence with score 5 or higher can be potentially included in the summary. Also note that in the rankings that humans assigned in the sentence-level study low numbers mean that a sentence is *not at all* suitable for a summary, so scores below 5 actually indicate that the sentences should not be included in a summary, while larger scores mean that sentences are *very suitable* for inclusion in a summary. We are trying to confirm the hypothesis that there are many sentences that have equal, overall low probability, of appearing in a summary. So in the sentence data these would be sentences that appear with a score in the range around 5 to 8, while the most important sentences will be with scores 9 or 10. With these differences in mind, we proceed to our study.

Each of the sentences in 20 multi-document sets was ranked, by three assessors. To obtain a single judgment for each sentence, we took the rounded average from the three judges. The percentage of judgments for each level are shown in table 2.4. On average across the 20 sets, 80% of the sentences were ranked between 8 and 5, while only 5% of the sentences had a score of 10 or 9, in comformity with our hypothesis. As with the study on the word level, few sentences have very high probability of being included in a summary, while a large percentage of the sentences have about equal, relatively lower probability of

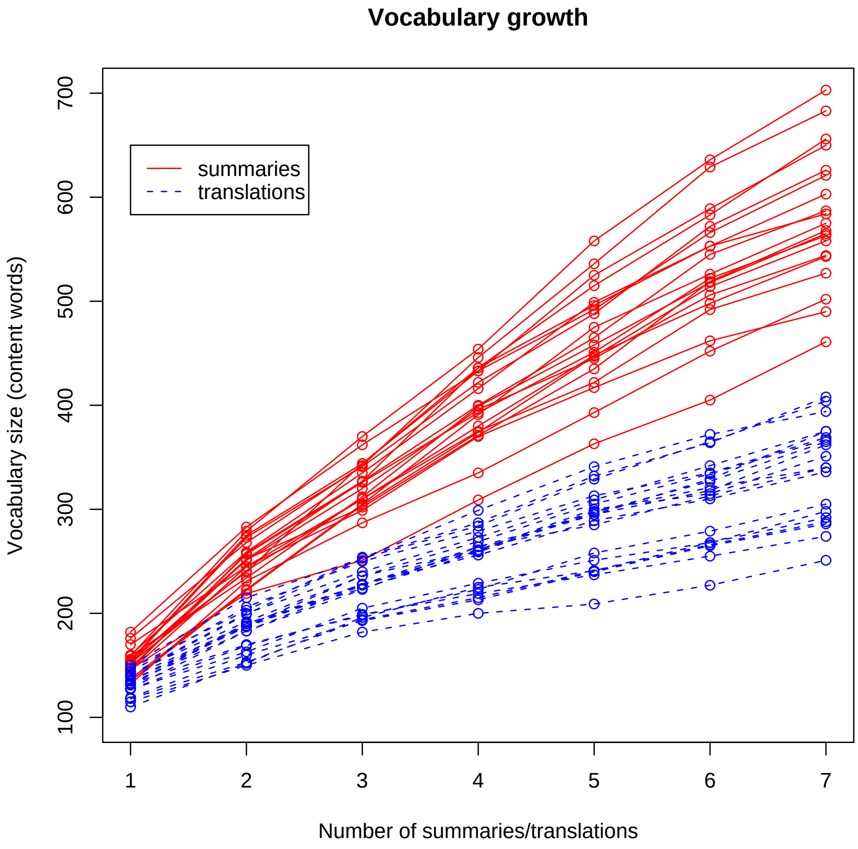

Figure 2.1: The growth of vocabulary with the addition of a new summary or translation. In summarization, each new summary brings much more new vocabulary items than a translation in the collection of translations. This confirms that the variation in summaries is bigger than one could explain with lexical variation alone and must be due to semantic variation.

10	9	8	7	6	5	4	3	2	1	0
0.14%	4.68%	17.90%	26.93%	23.58%	11.46%	7.74%	4.26%	2.27%	0.98%	0.20%

Table 2.7: Percentage of sentences in each category of appropriateness for inclusion in a summary

being included in a summary and can be chosen to form different, equally "valid" summaries.

Our study of multiple summaries of the same text confirms the intuition that content (words) is characterized by large number of rare events. The same observation about the content appropriate for inclusion in a summary can be made on the sentence level as well—the majority of sentences have a small chance of being included in a summary. It is thus reasonable to explore probabilistic models of summarization and we will turn to this problem in chapter 4, showing that the approach indeed leads to good results. It is also interesting to investigate the question of whether semantically defined content units, rather than words, would exhibit the same distribution properties across multiple summaries. In chapter 3 we will define a procedure for manual annotation of multiple summaries for semantically equivalent units and we will see that they exhibit distribution properties exactly analogous to those of content words and of sentences.

2.5 Related work

Task-based evaluations of summarization

We started this chapter with the description of a recent task-based evaluation of multi-document summarization that showed that this area of research can lead to improved user experience in a news-browsing setting. In the past, several task-based evaluations have been performed to establish the effectiveness of summarization systems, especially of single-document ones. In the (pre-DUC) TIPSTER Text Summarization Evaluation (SUMMAC), single-document summarization systems were evaluated in a task-based scenario developed around the tasks of real intelligence analysts (MKH^{+}02). This large-scale study compared the performance of a human in judging if a particular document is relevant to a topic of interest, by reading either the full document or a summary thereof. It established that automatic text summarization is very effective in relevance assessment tasks on news articles. Summaries as short as 17% of full text length sped up decision-making by almost a factor of 2 with no statistically significant degradation in accuracy. More recently, similar studies on the benefits of summarization for relevance judgments was performed by Dorr *et al.*, who propose a new metric for task-based evaluation, *Relevance-Prediction.*

Numerous studies have also been performed to investigate and confirm the usefulness of single document summaries for improvement of other automated tasks. For example, Sakai and Sparck Jones (SSJ01) present the most recent and extensive study (others include (BMR95) and several studies conducted in Japan and published in Japanese) on the usefulness of generic summaries for indexing in information retrieval. They show that indeed indexing for retrieval based on automatic summaries rather than full document text helps in certain scenarios for precision-oriented search. Another very interesting application of summarization for improvement of an automatic task has been reported by Burstein and Marcu (BM00). In their study, they examined the impact of summarization on the automatic topic classification module that is part of a system for automatic scoring of student GMAT essays. Their results show that summarization of the student essay significantly improves the performance of the topical analysis component. The conjectured reason for the improvement is that the students write these essays under time constraints and do not have sufficient time for revision and thus their writing contains some digressions and repetitions, which are removed by the summarization module, allowing for better assessment of the overall topic of the essay.

Evaluation analyses

In this chapter, we presented an analysis of results from large-scale evaluations of summarization in order to compare progress in the fields of multi- and single-document summarization, as well as to identify areas that need most improvement. The need to use evaluations as a means to further progress in the field has been especially emphasized with the beginning of DUC evaluations (MG01). The Marcu and Gerber study was specifically aimed at discovering which NLP modules are most useful for a multi-document summarizer so that they could focus on that during their preparation of the first DUC participation.

They were unable to satisfactorily answer the question, primarily due to the fact that there was very little data available for development before 2001. But after the first edition of DUC, much more interesting analyses could be carried out. McKeown *et al.* (MBE$^+$01) did an extensive analysis of the factors affecting evaluation results. They fitted an analysis of variance model and report that the variability of system scores was affected by four distinct factors, namely the document set (the input), (the human who constructs the) model summary, the size of the target summary and the peer summarizer. Moreover, they show that the human who constructed the model summary and the input document set had a larger effect on the outcome score than the peer system. In our study (Nen05), parts of which were presented here in section 2.2, we discussed similar findings that suggest future research directions. For example, to answer the need of diminishing the effect that the model summary has on evaluation, new evaluation methods based on multiple models have been proposed (Lin04; TvH04b; NP04) and we will discuss our particular proposal in the next chapter. The effect of the input to a summarizer on the ultimate summarizer performance has potential important consequences, since it indicates that some types of input can be handled more successfully by automatic systems than others, and understanding better the input characteristics can help decide which input should be directed to which summarizers, if several summarizers with different strengths are available (as in (MBE$^+$01)).

Studying human summarization performance

Several different aspects of human summarization performance have been studied in the past to inspire and guide the development of automatic summarization. Notably in the context of single document summarization, the cut-and-paste approach has been developed to mimic humans who edit extracted sentences using reduction to remove inessential phrases and combination to merge resulting phrases together as coherent sentences (JM99; Jin00; JM00). Central to developing the approach was the alignment of human abstracts with the original input documents in order to identify which input sentences were used as the bases of sentences in the abstract, and also to study the revision operations applied by the summary authors. The value for future research that the development of a corpus of aligned abstract and input document pairs and their texts has been recognized and several superior alignment methods have been proposed (BE03; DM04).

Banko and Vanderewende (BV04) used Jing's ideas as a foundation for their comparative study between the degree of extractiveness characteristic for single- and multi-document summarization. Jing's cut-and-paste approach for single document summarization suggests that larger text segments from the input documents can be combined to produce fluent summaries. Banko and Vanderwende's experiment aiming to discover if the same claim can be made about multi-document summaries as well, was based on the idea of "tiling" single- and multi-document human abstracts with tiles consisting of n-grams derived from the input provided to the summarizer. On average, the length of a tile for single-document summaries was 4.47, compared with 2.33 for multi-document summarization. They discovered that 61 out of all 1667 hand-written single-document summary sentences exactly matched a sentence in the source document, while there were no sentences in the multi-document summaries for which this was the case. In addition, 93% of the single-document summary sentences could be tiled using n-grams *coming from a single input sentence*, while none of the multi-document summary sentences could be tiled in this manner. Their study brings forward the fact that humans use a greater amount of generation and text reformulation when performing multi-document summarization task compared with the single-document summarization task, suggesting that summarization approaches that are very effective for single-document summarization might not perform as well in the multi-document setting. The study thus exemplifies the need to study human performance in order to derive specifications or guidelines for system development.

2.6 Conclusions

In this chapter, we analysed the characteristics of generic multi-document summarization. We showed that it is useful in the context of news aggregation and browsing system. We also studied the linguistic quality of multi-document summaries, and the differences in human content selection choices in summarization: both of these characteristics showed the need for more work in these areas. For linguistic quality, we observed that automatic summaries suffer from overall poor referential clarity, focus and coherence. In later chapters of the thesis, chapters 4 and 5 we will propose solutions to these problems through summary rewrite. For differences in human choices, we demonstrated that there are more differences in multi-document than single-document summarization and that these differences are larger than what could be explained by language variation alone. In the later chapters of this thesis, we will address this aspect of summarization by proposing an annotation scheme that allows us to analyse semantic similarities and differences between summaries using a better granularity than just content words, and will also propose an evaluation method that incorporates this observed variability into a reliable evaluation metric.

Chapter 3

The pyramid annotation scheme

In chapter 2, we performed a detailed comparison between the content in multiple human summaries on two levels of granularity—that of words and that of sentences, the latter based upon the relative utility studies performed by Radev and colleagues (RT03; RJB00). At both levels we observed a Zipfian distribution of content units (words and sentences respectively), with few units having a very high weight for inclusion in the summary, and many with only a small weight. [1] Nonetheless, the analysis is not fully satisfactory, because a linguistically motivated understanding of a semantic unit will be naturally larger than a content word and smaller than the complex sentences characteristic of newspaper writing (vD77; Hut87).

In this chapter we will introduce the pyramid annotation scheme—an annotation procedure specifically designed for comparative analysis of the content of several texts. In section 3.1, we introduce the annotation scheme and guidelines, describe an evaluation metric based on the annotation and discuss how pyramid analysis can serve as the basis for an empirically-motivated evaluation method and some of the benefits this method can offer. In section 3.2, we define possible alternative evaluation metrics based on the pyramid annotation, their properties and relative advantages (such as easier annotation requirements). Finally, in section 3.3, we discuss the large scale application of the pyramid evaluation method in the

[1] For words the weight was empirically determined as the number of the summarizers who used the content word in their summary, while for sentences the judgment of importance was directly obtained by asking an annotator to provide their subjective judgment.

2005 Document Understanding Conference, including a discussion on annotation reliability; we provide an analysis of semantically motivated content units, which complements the study of the distribution of importance of content units. In section 3.4, we discuss related work in examining the issues in text content comparison and summarization evaluation.

3.1 Annotation guidelines and pyramid scores

Our analysis of summary content is based on Summarization Content Units, or SCUs and we now proceed to define the concept. SCUs are semantically motivated units similar in spirit to the automatically identified elementary discourse units (Mar00; SM03), the manually marked information nuggets (Voo04) and factoids (vHT03; TvH04b), all of which are discussed in greater length in section 3.4. SCUs emerge from annotation of a collection of human summaries for the same input and are not bigger than a sentential clause. Rather than attempting to provide a semantic or functional characterization of what an SCU is, our annotation procedure defines how to compare summaries to locate the same or different SCUs.

The following example of the emergence of two SCUs is taken from a DUC 2003 test set. The sentences are indexed by a letter and number combination, the letter showing which summary the sentence came from and the number indicating the position of the sentence within its respective summary.

A1 In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.

B1 Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.

C1 Two Libyans, indicted for the bombing of a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.

D2 Two Libyan suspects were indicted in 1991.

The annotation starts with identifying similar sentences, like the four above, and then proceeds with finer grained inspection that can lead to identifying more tightly related subparts. We obtain two SCUs from the underlined portions of the sentences above. Each

SCU has a weight corresponding to the number of summaries it appears in; SCU1 has weight=4 and SCU2 has weight=3. The grammatical constituents contributing to an SCU are bracketed and coindexed with the SCU ID.

SCU1 (w=4): *two Libyans were officially accused of the Lockerbie bombing*

A1 [two Libyans]1 [indicted]1

B1 [Two Libyans were indicted]1

C1 [Two Libyans,]1 [indicted]1

D2 [Two Libyan suspects were indicted]1

SCU2 (w=3): *the indictment of the two Lockerbie suspects was in 1991*

A1 [in 1991]2

B1 [in 1991]2

D2 [in 1991.]2

The remaining parts of the four sentences above end up as contributors to nine different SCUs of different weight and granularity.

An SCU consists of a set of contributors that, in their sentential contexts, express the same semantic content. In addition, an SCU has a unique index, a weight, and a natural language label. The label, which is subject to revision throughout the annotation process, has three functions. First, it frees the annotation process from dependence on a semantic representation language. Second, it requires the annotator to be conscious of a specific *meaning* shared by all contributors. Third, because the contributors to an SCU are taken out of context, the label serves as a *reminder* of the full in-context meaning, as in the case of SCU2 above where the temporal PPs are about a specific event, the time of the indictment.

The complete set of annotation instructions is given in Appendix A. The instructions have evolved over the years, now addressing questions raised by annotators who applied the method in DUC 2005 as described in the following section. The initial guidelines were made publicly available in our Columbia University technical report (PN03).

After the annotation procedure is completed, the final SCUs can be partitioned in a pyramid. The partition is based on the weight of the SCU; each tier contains all and only the SCUs with the same weight. When we use annotations from four summaries, the pyramid will contain four tiers. SCUs of weight 4 are placed in the top tier and SCUs of

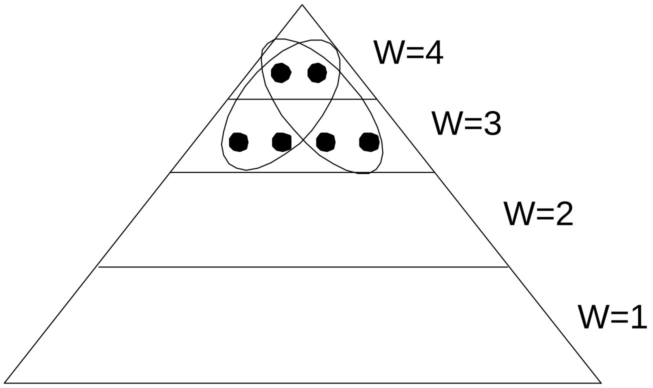

Figure 3.1: Two of six optimal summaries with 4 SCUs

weight 1 on the bottom, reflecting the fact that fewer SCUs are expressed in all summaries, more in three, and so on. For the mid-range tiers, neighboring tiers sometimes have the same number of SCUs. In descending tiers, SCUs become less important informationally since they emerged from fewer summaries. A more detailed discussion of these properties of SCU distribution will be presented in section 3.3.1.

We use the term "pyramid of order n" to refer to a pyramid with n tiers. Given a pyramid of order n, we can predict the optimal summary content—it should contain all the SCUs from the top tier, if length permits, SCUs from the next tier and so on. In short, in terms of maximizing information content value, an SCU from tier $(n-1)$ should not be expressed if all the SCUs in tier n have not been expressed. This characterization of optimal content ignores many complicating factors such as constraints for ordering SCU in the summary. However, we explicitly aim at developing a metric for evaluating *content selection*, under the assumption that a separate *linguistic quality* evaluation of the summaries will be done as well. The proposed characterization of optimal content is predictive: among summaries produced by humans, many seem equally good without having identical content. Figure 3.1, with two SCUs in the uppermost tier and four in the next, illustrates two of six optimal summaries of size four (in SCUs) that this pyramid predicts.

Based on a content pyramid, the informativeness of a new summary can be computed as the ratio of the sum of the weights of its SCUs to the sum of the weights of an optimal summary with the same number of SCUs. Such scores range from 0 to 1, with higher scores indicating that relatively more of the content is as highly weighted as possible. Like this, information that is included in more human summaries is awarded higher weight and importance. This decison assumes that the summary rewiters are equally capabale, and good at the summarization task. If this were not the case, information in the summary of more able summarizers can be awarded higher weight, for example.

We now present a precise formula to compute a score for a summary capturing the above intuitions about informativeness. Suppose the pyramid has n tiers, T_i, with tier T_n on top and T_1 on the bottom. The weight of SCUs in tier T_i will be i. There are alternative ways to assign the weights and the method does not depend on the specific weights assigned: the weight assignment we adopted is simply the most natural and intuitive one. Let $|T_i|$ denote the number of SCUs in tier T_i. Let D_i be the number of SCUs in the summary that appear in T_i. SCUs in a summary that do not appear in the pyramid are assigned weight zero. The total SCU weight $\mathcal{D}$ is:

$\mathcal{D} = \sum_{i=1}^{n} i \times D_i$

The optimal content score for a summary with X SCUs is:

$$\text{Max} = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|)$$

$$\text{where } j = \max_i (\sum_{t=i}^{n} |T_t| \geq X) \tag{3.1}$$

In the equation above, j is equal to the index of the lowest tier an optimally informative summary will draw from. This tier is the first one top down such that the sum of its cardinality and the cardinalities of tiers above it is greater than or equal to X (summary size in SCUs). For example, if X, is less than the cardinality of the most highly weighted tier, then $j = n$ and Max is simply $X \times n$ (the product of X and the highest weighting factor).

Then the pyramid score $\mathcal{P}$ is the ratio of D to Max. Because $\mathcal{P}$ compares the actual distribution of SCUs to an empirically determined weighting, it provides a direct comparison to the way human summarizers select information from source texts.

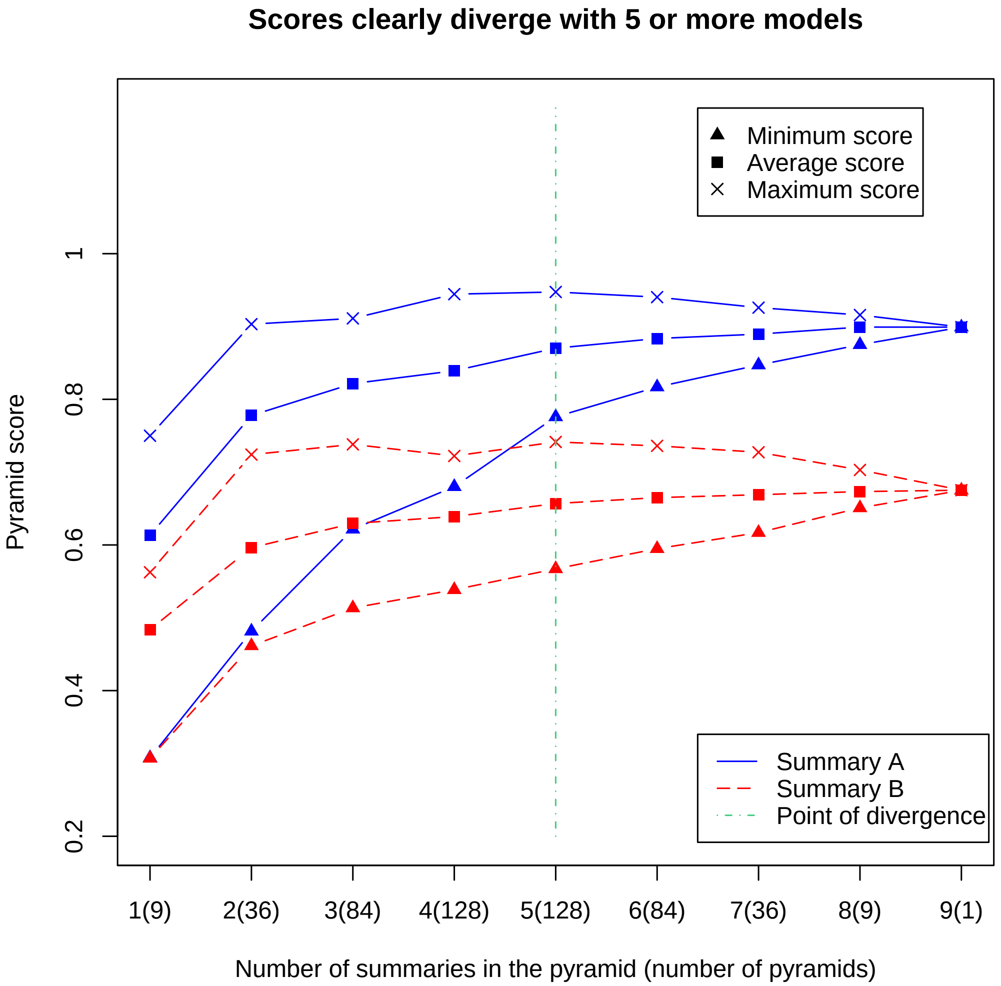

Figure 3.2: Minimum, maximum and average scores for two summaries summaries for pyramids of different size. Summary A is better than summary B as can be seen from the scores for pyramids of size 9, but with few models in the pyramid, it can be assigned scores that are lower than that for summary B.

3.1.1 Analysis of human summaries: behavior of scores as pyramid grows

Here we use three DUC 2003 summary sets for which four human summaries were written. In order to provide as broad a comparison as possible for the least annotation effort, we selected the set that received the highest DUC scores (D30042: Lockerbie), and the two that received the lowest (D31041: PAL; D31050: China). For each set, we collected six new summaries from advanced undergraduate and graduate students with evidence of superior verbal skills; we gave them the same instructions used by NIST to produce model summaries. This turned out to be a large enough corpus to investigate how many summaries a pyramid needs for score stability. We present results demonstrating the need for at least five summaries per pyramid, given this corpus of 100-word summaries. The two specific questions we examine in relation to the fact that summary scores change with the increase of the number of summaries in the pyramid are:

1. *How does variability of scores change as pyramid order increases?*
2. *At what order pyramid do scores become reliable?*

To have confidence in relative ranking of summaries by pyramid scores, we need to answer the above questions.

As we discussed in chapter 2, different people choose different content for inclusion in their summaries, so a summary under evaluation could receive a rather different score depending on which summary is chosen to be the model.[2] We observe that with only a few summaries in a pyramid, there is insufficient data for the scores associated with a pyramid generated from one combination of a few summaries to be relatively the same as those using a different combination of a few summaries. Empirically, we observed that as pyramids grow larger, and the range between higher weight and lower weight SCUs grows larger, scores stabilize. This makes sense in light of the fact that a score is dominated by the higher weight SCUs that appear in a summary. However, we wanted to study more precisely at what point scores become independent of the choice of models that populate the pyramid. We conducted three experiments to locate the point at which scores stabilize across our

[2] At the end of chapter 2 we mentioned the work of McKeown *et al.* that showed that in evaluations based on a single model, the choice of the model had a significant impact on the scores assigned to summaries.

three datasets. Each experiment supports the same conclusion (that about five summaries are needed), thus reinforcing the validity of the result.

Our first step in investigating score variability was to examine all pairs of summaries where the difference in scores for an order 9 pyramid was greater than 0.1; there were 68 such pairs out of 135 total. All such pairs exhibit the same pattern illustrated in Figure 3.2 for two summaries we call 'Summary A' (shown in blue) and 'Summary B' (shown in red). The x-axis on the plot shows how many summaries were used in the pyramid (and in brackets, the number of pyramids of that size that could be constructed with the nine available model summaries) and the y-axis shows the minimum (marked on the plot by a triangle), maximum (marked by a cross) and average (marked by a square) scores for each of the summaries for a given order of pyramid.[3] Of the two, 'Summary A' has the higher score for the order 9 pyramid, and is perceivably more informative. Averaging over all order-1 pyramids, the score of 'Summary A' is higher than that for 'Summary B' (with all orders of pyramids, including that for order-1, the blue square representing the average score for 'Summary A' across all possible pyramids is above the red square that represents the average score for 'Summary B'). But some individual order-1 pyramids might yield a higher score for 'Summary B': the minimum score assigned by some pyramid to 'Summary A' (blue triangle) is lower than the average score for the worse summary B. The score variability at order-1 is huge: it can be as high as 0.5, with scores for summary A varying between around 0.3 to close to 0.8. With higher order pyramids, scores stabilize: the difference between the maximum and minimum score each summary could be assigned diminishes, and even the lowest score assigned to the better summary (A) is higher than any score for the worse summary (B). Specifically, in our data, if summaries *diverge* at some point as in Figure 3.2, meaning that the minimum score for the better summary is higher than the maximum score for the worse summary, the size of the divergence never decreases as pyramid order increases. This is visually expressed in the figure by the growing distance between the blue triangles and the red crosses. The vertical green dotted line at pyramids of order 5 marks the first occurrence of divergence in the graph. For pyramids of order > 4, 'Summary A' and 'Summary B' never receive scores that would reverse their ranking,

[3] Note that we connected data points with lines to make the graph more readable.

regardless which model summaries are used in the pyramids.

For all pairs of divergent summaries, the relationship of scores follows the same pattern we see in Figure 3.2 and the point of divergence where the scores for one summary become consistently higher than those of the other, was found to be stable—in all pair instances, if summary A gets higher scores than summary B for all pyramids of order n, than A gets higher scores for pyramids of order $\geq n$. We analyzed the score distributions for all 67 pairs of summaries the order-9 scores for which differed by more than 0.1, in order to determine what order of pyramid is required to reliably determine which is the better one. The expected value for the point of divergence of scores, in terms of number of summaries in the pyramid, is 5.5.

We take the scores assigned at order 9 pyramids as being a reliable metric on the assumption that the pattern we have observed in our data is a general one, namely that variance always decreases with increasing orders of pyramid, and that once divergence of scores occurs, the better summary never gets lower score than the worse for any model of higher order.

We postulate that summaries whose scores differ by less than 0.06 have roughly the same informativeness. The assumption is supported by two facts. First, this corresponds to the difference in scores for the same summary when the pyramid annotation has been performed by two independent annotators (see (NP04) for details). In later studies in the context of DUC 2005, it was also shown that scores based on peer annotations produced by novice annotators given the same pyramid also differ on average by 0.06 (PNMS05). Second, the pairs of summaries whose scores never clearly diverged, had scores differing by less than 0.06 at pyramid order 9. So we assume that differences in score by less than 0.06 do not translate to meaningful differences in information quality and proceed to examine how the relative difference between two summaries at order-9 pyramids could change if we used pyramids of lower order instead.

Now, for each pair of summaries $(sum1, sum2)$, we can say whether they are roughly the same when evaluated against a pyramid of order n and we will denote this as $|sum1| ==_n |sum2|$, (scores differ by less than 0.06 for some pyramid of order n) or different (scores differ by more than 0.06 for all pyramids of order n) and we will use the notation $|sum1| <_n |sum2|$

if the score for $sum2$ is higher.

When pyramids of lower order are used, the following errors can happen, with the associated probabilities:

E$_1$: $|sum1| ==_9 |sum2|$ but $|sum1| <_n |sum2|$ or $|sum1| >_n |sum2|$ at some lower order n pyramid. The conditional probability of this type of error is $p_1 = P(|sum1| >_n |sum2|||sum1| ==_9 |sum2|)$. In this type of error, summaries that are essentially the same in terms of informativeness will be falsely deemed different if a pyramid of lower order is used.

E$_2$: $|sum1| <_9 |sum2|$ but at a lower order $|sum1| ==_n |sum2|$. This error corresponds to "losing ability to discern", and a pyramid with fewer models will not manifest a difference that can be detected if nine models were used. Here, $p_2 = P(|sum1| ==_n |sum2|||sum1| <_9 |sum2|)$.

E$_3$: $|sum1| <_9 |sum2|$ but at lower level $|sum1| >_n |sum2|$ Here, $p_3 = P(|sum1| >_n |sum2|||sum1| <_9 |sum2|) + P(|sum1| <_n |sum2||sum1| >_n |sum2|)$. This is the most severe kind of mistake and ideally it should never happen, with the better summary getting a much lower score than the worse one. Note that such error can happen only for models of order lower than their point of divergence.

Empirical estimates for the probabilities p_1, p_2 and p_3 can be computed directly by counting how many times the particular error occurs for all possible pyramids of order n. By taking each pyramid that does not contain either of $sum1$ or $sum2$ and comparing the scores they are assigned, the probabilities in Table 5 are obtained. We computed probabilities for pairs of summaries for the same set, then summed the counts for error occurrence across sets. The order of the pyramid is shown in the first column of the table, labeled n. The last column of the table, "Data points", shows how many pyramids of a given order were examined when computing the probabilities. The total probability of error $p = p1 * P(|sum1| ==_9 |sum2|) + (p2 + p3) * (1 - P(|sum1| ==_9 |sum2|))$ is also shown in Table 5.

Table 5 shows that for order-4 pyramids, the errors of type E$_3$ are ruled out. At order-5 pyramids, the total probability of error drops to 0.1 and is mainly due to error E$_2$, which is the mildest one.

Choosing a desirable order of pyramid involves balancing the two desiderata of having less data to annotate and score stability. Our data suggest that for this corpus, 4 or 5 summaries provide an optimal balance of annotation effort with score stability. This is reconfirmed by our following analysis of ranking stability.

n	p1	p2	p3	p	data points
1	0.41	0.23	0.08	0.35	1080
2	0.27	0.23	0.03	0.26	3780
3	0.16	0.19	0.01	0.18	7560
4	0.09	0.17	0.00	0.14	9550
5	0.05	0.14	0.00	0.10	7560
6	0.02	0.10	0.00	0.06	3780
7	0.01	0.06	0.00	0.04	1080
8	0.00	0.01	0.00	0.01	135

Table 3.1: Probabilities of errors E1, E2, E3 (p_1, p_2 and p_3 respectively, and total probability of error (p). The first column shows the order of the pyramid, equal to the number of model summaries it is constructed from. The last column gives the number of observations used to compute the probabilities.

In order to study the issue of how the pyramid scores behave when several summarizers are compared, not just two, for each set we randomly selected 5 peer summaries and constructed pyramids consisting of all possible subsets of the remaining five. We computed the Spearman rank-correlation coefficient for the ranking of the 5 peer summaries compared to the ranking of the same summaries given by the order-9 pyramid. Spearman coefficient r_s (DM69) ranges from -1 to 1, and the sign of the coefficient shows whether the two rankings are correlated negatively or positively and its absolute value shows the strength of the correlation. The statistic r_s can be used to test the hypothesis that the two ways to assign scores leading to the respective rankings are independent. The null hypothesis can be rejected with one-sided test with level of significance $\alpha = 0.05$, given our sample size $N = 5$, if $r_s \geq 0.85$.

Since there are multiple pyramids of order $n \leq 5$, we computed the average ranking coefficient, as shown in Table 3.2. Again we can see that in order to have a ranking of the summaries that is reasonably close to the rankings produced by a pyramid of order $n = 9$, 4 or more summaries should be used.

n	average r_s	# pyramids
1	0.41	15
2	0.65	30
3	0.77	30
4	0.87	15
5	1.00	3

Table 3.2: Spearman correlation coefficient average for pyramids of order $n \leq 5$

3.2 Other scores based on pyramid annotation

The pyramid score defined and studied in the previous section represents only one of the possible ways for incorporating the content unit analysis into a score reflecting the appropriateness of a content in a summary. The pyramid score is similar to a precision metric—it reflects how many of the content units that were included are as highly weighted as possible and it penalizes the use of a content unit when a more highly weighted one is available and not used.

Alternatively, one can imagine a pyramid score corresponding to recall. Such recall oriented score could be defined, for example, as the weight of the content units in the summary normalized by the weight of an ideally informative summary of SCU size equal to the average SCU size the human summaries in the pyramid. So again, the score would be the ratio between $\mathcal{D}$ (the sum of weights of SCUs expressed in the summary) and Max (the optimal score of a summary of size X, defined in formula 3.1), as defined in section 3.1, but this time X will not be the SCU length of the evaluated peer, but rather the average number of SCUs in the model summaries used for the creation of the pyramid, X_a:

$$X_a = \frac{\sum_{i=1}^{n} i \times |T_i|}{n} \tag{3.2}$$

This score, which we will call *modified pyramid score*, measures if the summary under evaluation is as informative as one would expect given the human models. If a summary covers some of the highly weighted content units, and also contains multiple content units that do not appear in the model pyramid, such summary would have a high modified pyramid score and possibly quite low original score when the summary contains more content units than the average human model. In the next section 3.3, we will discuss the findings from the application of the pyramid evaluation method in DUC 2005 where the modified pyramid score showed better qualities than the original pyramid scores—it proved to be less sensitive to peer annotation errors, it distinguished better between systems and had higher correlation with other manual evaluation metrics such as responsiveness judgments.

Another possibility for a score can ignore the weighting of content units altogether. The pyramid annotation can be used simply to obtain a pool of content units that are likely to appear in the summary, similarly to the way *nuggets* are used for evaluation of question-answering systems (Voo04). In this scenario, the standard precision and recall used in information retrieval can be computed. Earlier we defined D_i as the number of SCUs in a summary under evaluation that appear in tier T_i of the pyramid. In particular, D_0 is the number of SCUs in the peer that do not appear in the pyramid. Then we can straightforwardly define

$$Recall = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} T_i} \tag{3.3}$$

and

$$Precsion = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=0}^{n} D_i} \tag{3.4}$$

Recall is equal to the fraction of content units in the peer summary that are also in the pyramid and precision is the ratio of SCUs from the pyramid that are expressed in the peer to all SCUs expressed in the peer. Such scores would not incorporate all the knowledge derivable from SCU analysis of multiple summaries—the benefit from the use of multiple models will be only that a bigger pool of potential content units can be collected. But the importance weighting will not be used. A notable difference of the precision/recall approach proposed here and that used in evaluation of question answering systems is that the pyramid method one is based on an analysis of *human models*, while the information nuggets in question-answering evaluation are obtained by analyzing (mostly) the output of automatic systems, thus making it impossible to claim that an occurrence in the answer provides an empirical evidence for the importance of the nugget.

Another interesting possibility is to use each summary content unit as the unit of evaluation. Such approach is similar to the one used in machine translation, where human judgments for quality are collected on sentence by sentence basis, rather than for a complete text. Such score can be used when the goal is to compare systems and will work in the following way. Say there are N content units in a pyramid, $N = \sum_{i=1}^{n} T_i$. Each peer summary will be associated with a binary vector S and S[k] = 1 if the kth content unit from the pyramid is expressed in the peer (and is 0 otherwise). Thus, when comparing two summaries from different systems, for each test set we will get a vector of observations, rather than a single number as the original or modified pyramid scores do. This means that one could apply a paired statistical test (a t-test for example) to the two vectors and test if *two different summaries for the same set* are *statistically* significantly different. It is not possible to make conclusions of this sort from the original pyramid scores because a vector of observations is necessary to compute the variance of the scores. Similarly, when comparing two systems across a full test set, the content unit based evaluation would give more data points which can be used to compare the systems. Say there are Y test sets. If the per summary scores are used, the basis for comparison between the two systems will consist of a vector of Y observations. But there will be about $Y * N_a$ data points for content unit based evaluation, where N_a is the expected number of SCUs in a pyramid. For the three sets that we described earlier in this chapter, there were 36 SCUs in an average pyramid. So if we want to compare two systems, we will have 3 data points if pyramid scores are used and $36 * 3 = 114$ data points if the evaluation is done on the basis of content units. Such gain in data can make possible reaching statistically significant conclusions even when few test sets are available. In the experiments for DUC 2005, which we will describe next, there were 20 test sets for pyramid evaluation. We experimented with the content unit based score, and indeed saw several more significant differences between system compared to the modified pyramid score. So the introduction of the new metric seemed unnecessary in the DUC context, with only few gains in significance. But in other settings, where fewer test sets are available, the per content unit evaluation can be very helpful.

3.3 Pyramid evaluation at DUC 2005

In the 2005 Document Understanding Conference, special attention was devoted to the study of evaluation metrics. The pyramid semantic-centered evaluation was one of the metrics used. 20 test sets were evaluated with the pyramid method, in addition to the linguistic quality evaluation discussed in section 2, responsiveness and the automatic ROUGE metrics. The task in that edition of the conference was to produce a 250-word summary in response to a question topic. The responsiveness of each summary was judged by a human on a scale from 1 (least responsive) to 5 (most responsive), with the additional restriction that each category should be assigned to at least one of the summaries.

An example of a topic for which the systems needed to find an answer is shown below:

```
<topic>
<num> d311i </num>
<title> VW/GM Industrial Espionage </title>

<narr>
Explain the industrial espionage case involving VW and GM. Identify
the issues, charges, people, and government involvement. Report the
progress and resolution of the case. Include any other relevant
factors or effects of the case on the industry.
</narr>
```

Pyramid evaluation was applied to 27 peers for each of the 20 test sets. The peers included one baseline (the first 250 words of the latest document), 24 automatic peers and two human peers. The model pyramid for each set was constructed based on seven human-authored summaries. The 20 pyramids were constructed by a team at Columbia University, and the peer annotation was performed by DUC participants and additional volunteers who were interested in the pyramid annotation. There were total of 26 peer annotators, which allowed the freedom to have the same set of summaries annotated by two different annotators and study the annotation reliability. Details on the DUC 2005 pyramid evaluation beyond those presented in the chapter can be found in (PNMS05).

The evaluation allowed for two kinds of summarization-related studies:

1. The model pyramids allow better understanding of the distribution of information in a pool of human summaries, measured in semantically defined content units. The study confirms the hypothesis suggested by the analysis based on the distribution of content words and sentences that was described in chapter 2—-content units follow a Zipfian distribution characteristic for complex optimization problems.

2. The evaluation results allows for comparison between the semantic-based pyramid evaluation and the other metrics used in the conference.

The following sections offer an indepth discussion of these two topics concerning the pyramid annotation.

3.3.1 Zipfian distribution of content units

In chapter 2 we hypothesized that content units from the input are probabilistically emitted in a summary (see the beginning of section 2.4) and that there are only a few content units with high probability of being emitted in the summary, and a very large group of units with quite low probability. Essentially, the distribution of the importance (probability) of content units is Zipfian—figure 3.3 shows the distribution of each importance class, on three levels of granularity, Summarization Content Units, words and sentences. At all levels, the initial hypothesis of Zipfian distribution in the importance classes is confirmed. Remember that for the purpose of our study, importance is equated to the number of humans that uses a content unit in a summary.

In all three plots, the x axis shows the importance weight of the unit and the y axis shows how many units of the given importance class appeared in our corpus. The SCU and content word counts are from the DUC 2005 summary corpus consisting of 20 input sets and seven human summaries for these clusters. The importance classes of SCUs/words are simply the number of human summarizers who included the given SCU/word in their summary. Thus the bar over "class 7" on the x axis represents the number of SCUs in the 20 sets that had weight or importance equal to 7. The bottom graph on the importance of sentences comes from direct judgment on the importance of each sentence done by the relative utility method discussed in chapter 2 and further described in the related work section of this chapter. In the relative utility approach, each sentence in the input was judged on a scale from 10 to 0 with respect to its appropriateness for inclusion in the summary, where 10 means the sentence should definitely be included in the summary and 0 means the sentence should definitely not be included in the summary. For our purposes, all sentences that were assigned a relative utility judgment strictly less than 7 are collapsed together. The corpus for relative utility judgments is *different* from the DUC 2005 corpus used for the other two granularities.

An inspection of 3.3 reveals a striking similarity between the importance class distributions in terms of words and SCUs. The observed similarity in the distributions gives some insight into explaining the fact that automatic intrinsic evaluation of content selection in summarization based on unigram overlap with a set of models have high correlation with semantic-based evaluation methods (LH03; Lin04).

Zipfian distribution is characteristic of complex problems involving the optimization of multiple constraints, for example that of city sizes, the distribution of wealth (Zip65), and the connectivity between webpages or the citation patterns in science (BA99). Indeed, the process of summarization itself is constrained by multiple cognitive factors such as prior knowledge and personal interest, as well as the properties of the input documents. In fact in the past couple of years several applications of graph-based algorithms originally developed for estimating importance of webpages have been applied to summarization, with very good results—(ER04a; MT04; VBM04). The researchers who proposed the graph-based approach conclude that graphs are a very powerful modeling tool. But here we have demonstrated that the good performance of the algorithms such as PageRank (PBMW98) that were originally developed for the web lead to good results in summarization because in fact content units in summaries have the same distribution as webpage connectivity. While from an engineering perspective it is not so important *why* an approach to solving a problem works well, having

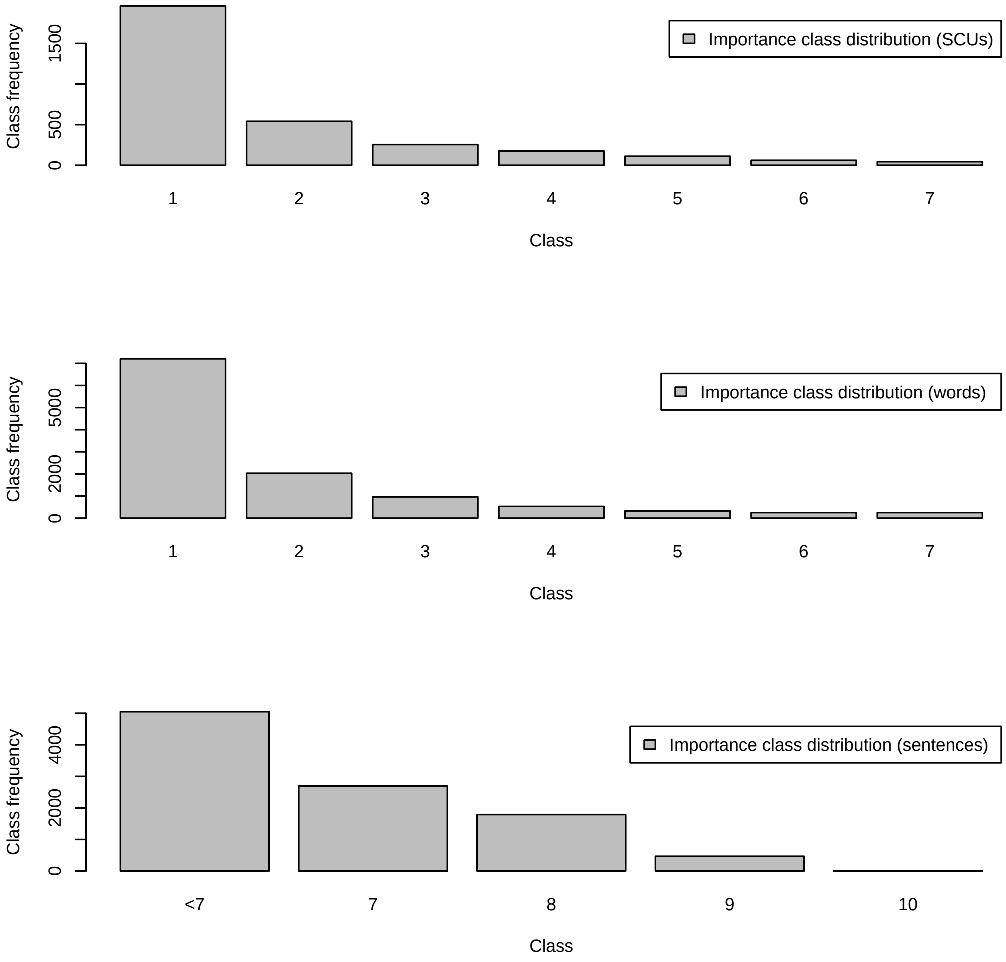

Figure 3.3: Size of "importance" classes defined in terms of content units, words and sentences. The shape of the distributions is the same across the different granularities, with few units of a very high weight and a numerous units of low weight. The importance of sentences is based on subjective human judgments, while the importance of content words and SCUs are empirically estimated from the summary content chosen by humans.

an insight into the whys surrounding the summarization process are very useful—they can lead to understanding deeper cognitive and linguistic aspects of summarization that can eventually feed back useful ideas for system development. For example, in the chapter on content selection (chapter 4) we will use a model of importance that relies on frequency information and it proved to be powerful for generic news summarization. It is a more direct model for example than the graph-based models that use word co-occurrence to build links between graph nodes, as well as to assign weight for these arcs (a more detailed comparison between the two approaches will be presented in the following chapter, 4). But the distributional observations presented here suggest that some of the more mathematically sophisticated "Large Number of Rare Events" models presented in (Baa01) could lead to even better results in content selection. We plan to experiment with some of these models in future work.

Obviously, the study of distribution of content units has implications for summarization evaluation. A meaningful metric should incorporate the findings, in a way similar to the pyramid or factoid evaluation methods. From the Zipfian distribution of content units we can expect that the probability of a content unit to appear in a summary cannot be estimated with convergence—as (Baa01) point out, the consideration of a larger number of human models, the overall (average) probability of content units will decrease. Such conclusion is confirmed by our empirical findings as well—in SCU pyramids constructed from four human summaries, the average weight of an SCU was 2.4 (DUC'03 data), while in pyramids constructed from seven model summaries the average SCU weight was 1.9 (DUC'05 data). What the pyramid method achieves is to obtain a *relative* weight for the content units. After 5 summaries are analyzed, one gets a good working estimate of which content units are *more important than others that can appear in a summary.* This relative importance is unlikely to change with the addition of more summaries and provides for a reliable intrinsic evaluation metric.

3.3.2 Annotation reliability

In order to study the reliability of peer evaluation using pyramid annotations, the 27 peers for six of the test sets were annotated by two different annotators, using the guidelines provided here in Appendix A. The annotators were encouraged to use th DUC 2005 mailing list to request further clarifications. Both annotators received the same model pyramid constructed from seven human summaries and each independently evaluated the 27 peers for that set. This means that each annotator had to read a peer summary and decide if the information in the peer is equivalent to some content unit in the model pyramid, and to identify the exact content unit. The annotators were encouraged to revise their annotations after looking at more peers and were provided with a script that ensured consistant annotation of the same text in different summaries.

There are two levels at which the annotator agreement could be measured:

Semantic At the semantic level, one can be interested if the two annotators can use the annotation guidelines to reliably make decisions of what content, potentially realized with different syntactic constructions and vocabulary items, can be considered equal semantically.

Scorewise In the context of summary evaluation, one can be interested exclusively in the scores that each peer receives, and not in the judgments that led to the score. Thus of interest will be if and by how much the score of the summary obtained under one annotator differs from that under the other annotator.

We now proceed with the study of both aspects of the annotation reliablily. In both respects the method is sound—the overall kappa statistic between the pairs of annotators reaches 0.65 when content units of weight 1 are excluded from consideration, and the average difference between the pairs of scores for each summary is less than 0.06. Detailed analysis follows.

Do annotators identifying content units reliably?

The question of measuring reliability of an annotation scheme is always an important one and has received considerable attention in the natural language processing community, especially since Carletta's influential paper (Car96) that specifically discussed the importance of reliability. The appropriate choice of statistics to measure reliability has been a matter of discussion, notably in (DG04; Kri04; CW05), the first suggesting that several agreement statistics should be reported in conjunction, while the latter two claiming only one should be chosen. Here we will report percent agreement and kappa, as recommended by di Eugenio and Glass (DG04) who suggests that these two statistics should be reported together, along with Cohen's kappa and that the three statistics in conjunction provide a good picture of agreement.

Percent agreement measures the proportion of agreement between coders and is the ratio between the number of content units coded in the same manner by both annotators as wither present or absent in the peer and the total number of content units in a pyramid. It does not include any correction for chance agreement. **Kappa** computes the chance-corrected agreement between annotators, with the chance agreement computed by assuming equal coder category distribution. In the peer annotation task, the annotator is given a master pyramid, annotated separately by an annotator and adjusted after a discussion with other annotators. The task in the peer evaluation is to take a new peer summary (that did not contribute to the identification of content units in the model pyramid) and then identify all content units from the pyramid that also appear in the peer summary. Effectively this is equal to the task of answering with *yes* or *no*, for each SCU in the pyramid, the question *"Does this SCU appear in the peer summary?"* In addition, the annotator also had to highlight the portion of the peer summary that expressed the SCU in case of an *yes* answer, but this aspect of the annotation does not contribute to the reliability study. The model pyramids contained on average 156 SCUs across the 20 sets, while an average peer covered about 10 SCUs from the pyramid. So the number of SCUs that will not be expressed in the peer will greatly outnumber the ones expressed. As a consequence, the percent agreement will be higher than the actual agreement between annotators and the kappa statistic will be lower, since a the estimated chance agreement for the majority category will be high. We report both: the kappa and percent agreement statistics for each of the six sets are shown in table 3.3.2. In particular the kappa statistic ranges from 0.46 to 0.62. These are somewhat low reliability figures, indicating the difficulty of the annotation task. In fact, we noticed that more and longer model summaries in DUC 2005 (250 words; 7 models) lead to longer annotation times. We computed the two reliability statistics while excluding the SCUs of weight one, which represented about half of the SCUs in a pyramid—the kappa and percent

Set	Kappa statistic	Percent agreement
set 324	0.62 (0.66)	0.90 (0.90)
set 400	0.50 (0.65)	0.96 (0.94)
set 407	0.46 (0.42)	0.95 (0.90)
set 426	0.60 (0.69)	0.94 (0.94)
set 633	0.55 (0.67)	0.96 (0.96)
set 695	0.57 (0.71)	0.96 (0.96)

Table 3.3: Kappa and percent agreement statistics for two independent annotators in six of the DUC 2005 peer annotation sets. For each set, 27 peers were evaluated. The numbers in brackets show the same statistics computed over annotations excluding the SCUs of weight 1.

agreement agreement of the reduced SCU set are shown in brackets in table 3.3.2. Indeed, kappa figures go up when computed with this reduced SCU set, close to or exceeding the generally recommended threshold of 0.67 for five of the six sets. This is an indication that while it might be desirable to get a larger number of model summaries in order to better estimate the weight of content units, the larger number leads to longer time spend doing the annotation, and makes the annotation more difficult, lowering the reliability.

Computed over all six sets annotated by two annotators by taking the scu annotation across all sets in one pool, we obtain a kappa statistic of 0.57 and percent agreement 0.95; without weight 1 SCUs the two statistics are 0.65 and 0.93 respectively. The general recommendation for a threshold at which an annotation can be considered reliable is kappa of 0.67, which is about what is achieved in pyramid annotation when weight 1 SCUs are excluded.

In conclusion, the SCU annotation task is difficult for novices: as (CW05) discuss, lower kappa statistics are indicative of the difficulty of the task the annotators were asked to perform. Better training can lead to higher agreement as measured by kappa as we discuss in our presentation of the factoid method in the related work section of this chapter.

Additional discussion on the annotation reliability in the DUC'05 exercise, including alpha statistics (Kri80), can be found in (PNMS05).

By how much do scores differ?

In the previous section we studied the *reliability* of peer pyramid annotation. The study showed that the task is difficult, but acceptable reliability can be achieved even when the annotation is performed by untrained annotators. Still, in terms of evaluation, the reliability of the annotation is not the most important property—rather, the stability of the scores that summaries receive based on the two annotations is of importance. A small difference would show that even if there are disagreements if the peer annotation, the score is not impacted. For the six sets that were annotated by two annotators and that we used for the reliability discussion in the previous section, we computed the difference between the scores resulting from the two different annotations. The average absolute value of the difference in scores across the 162 pairs of summaries annotated by two annotators was 0.0617 for the original pyramid score and 0.0555 for the modified pyramid score. These numbers are very reassuring because they are very close to the empirically estimated difference of scores that we postulated in 3.1.

Tables 3.4 shows the average paired difference for each set for the original (in column 2) and modified scores (in column 3) respectively. It also shows p-values for a parametric t-test (those for non-parametric Wilcoxon rank-sum test were also computed and were very similar to the p-values for the t-test and are not reported here for this reason) for the null hypothesis that the annotation of one annotator led to scores that are significantly higher than the scores based on the other's annotation.

The average differences in the six sets are overall in the expected range, smaller than 0.06. The only exception is set 324, where the scores differed on average by 0.1 for the modified score and 0.7 for the original pyramid scores. One of the annotators for this set 324 actually reported being pressed for time, and not using the script provided to annotators to ensure consistency. The fact the set in which this annotator was involved had the largest differences across annotator scores is indicative of the generally recognized need to doublecheck annotations before finalizing them. The results from the t-test also suggest that additional training of the novice annotators might be beneficial: they show that the differences in score are significant, that is, that one annotator consistently annotated in a way that lead to higher scores compared to the other annotator for the same set (maybe by

Set	Original score (t-test p-val)	Modified score (t-test p-val)
324	-0.0713 (0.0001)	-0.1048 (0.0001)
400	-0.0062 (0.6654)	-0.0401 (0.0002)
407	-0.0413 (0.0113)	-0.0401 (0.0002)
426	0.0142 (0.2096)	-0.0238 (0.0113)
633	0.0289 (0.0353)	0.0200 (0.1155)
695	-0.0506 (0.0001)	-0.0357 (0.0001)

Table 3.4: Average difference between the original and modified pyramid scores from two independent annotations. The p-values from a paired t-test are given in brackets.

being more liberal in equating content). For the original scores, two out of the six sets had significantly different scores (the exceptions were sets 400 and 426), while for the modified score, only the difference for set 633 is not significant. The stronger significance of differences for the modified score compared to the original score is consistent with the fact that many annotators reported that they were unsure how to annotate content in the peer that is not in the pyramid. Many annotators adopted their own strategy for annotating such content, leading to random changes in the original score: the differences there were not systematic. For the modified score, which does not take into account the annotation of content that does not appear in the pyramid, the differences are much more systematic, indicated by the lower p-values for each set. Annotator training, or a protocol for doublechecking the annotations, possibly by another annotator, are likely to further reduce the observed differences.

Overall the results are good: even with novice annotators who had not been previously exposed to the annotation procedure, the differences between scores were in the range expected based on our prior research. But there is still room for improvement which can be achieved by training and a post-annotation adjudication procedure that we plan to introduce for future evaluations.

Correlations with other evaluation metrics

In this section, we will overview the correlations between the manual and automatic metrics used in DUC. The study of correlations is important in order to identify which metrics are mutually redundant or substitutable. For example, if two metrics A and B have correlation exceeding 0.95, and if we know the scores for one metric, say A, then we can predict the scores for the other (B) with a very high accuracy. If the scores for metric B are more difficult to obtain than those for metric A, for example they require more annotation, more human subjects, etc, then we can say that the metrics are mutually substitutable and simply use metric A in place of metric B. This situation usually arises when one of the metrics is automatic (easier to produce) and the other is manual (more difficult and expensive to produce). In the case when scores for both metrics with high correlation above 0.95 are equally easy/difficult to produce, it is advisable to chose and report only one of them, since the other does not bring in any new information into the analyses. Likewise, if two metrics do not have perfect correlation, it means that they give information on some orthogonal qualities of the summaries and can be used jointly for overall assessment.

Table 3.3.2 shows the correlations between pyramid scores and the other official metrics from DUC 2005—responsiveness and bigram overlap (ROUGE-2) and unigram overlap and skip bigram (ROUGE-SU4). The responsiveness judgments were solicited by two NIST annotators (responsiveness-1 and responsiveness-2), who ranked all summaries for the same input on a scale from 1 to 5. The two ROUGE metrics are automatically computed by comparing a summary to a pool of human models on the basis of n-gram overlap. The numbers are computed using only the 25 automatic peers—when the humans are included, all correlations exceed 0.92. But in all metrics the human performance is much better and human and automatic summarizers form different clusters, so a more fair and realistic analysis of correlations includes only the automatic summarizers.

None of the correlations in the table are statistically significantly different from any other, showing that no pair of metrics is more similar to each other than the other pairs. They are all rather high and significantly different from zero. The two variants of the pyramid scores (original and modified) are very highly correlated, with Pearson's correlation coefficient of 0.96. The two automatic metrics are also very highly correlated (0.98),

	Pyr (mod)	Respons-1	Respons-2	ROUGE-2	ROUGE-SU4
Pyr (orig)	0.96	0.77	0.86	0.84	0.80
Pyr (mod)		0.81	0.90	0.90	0.86
Respons-1			0.83	0.92	0.92
Respons-2				0.88	0.87
ROUGE-2					0.98

Table 3.5: Pearson's correlation between the different evaluation metrics used in DUC 2005. Computed for 25 automatic peers over 20 test sets.

indicating that the metrics are mutually redundant. At the same time, the two sets of responsiveness judgments, given by two different judges under the same guidelines, have a correlations of only 0.83, confirming that the metric is subjective and different scores are likely to be assigned by different humans. The correlation between responsiveness-2 and the modified pyramid score is as high as 0.9 but still the metrics are not mutually redundant and each reveals information about the summary quality that is not captured by the other. The automatic metrics correlate quite well with the manual metrics, with Pearson's correlation in the range 0.80 to 0.92 but still do not seem to be high enough to suggest that the automatic metrics can be used to replace manual metrics, which is very comparable to results from previous years on multi-document test sets (Lin04).

In previous years, a manual evaluation protocol based on a comparison between a *single* model and a peer was used. In his studies, Lin compared the correlations between these manual scores and several versions of automatic scores. Very good results were achieved for single document summarization and for very short summaries of 10 words where the correlation between the automatic and manual metrics was 0.99 and 0.98 respectively. But for 100-word multi-document summaries, the best correlation between an automatic metric and the manual metric was 0.81: the correlations for multi-document summarization are not as high as the ones achieved in automatic evaluation metrics for machine translations and for other summarization tasks, where Pearson's correlations between manual and automatic scores was close to perfect 0.99 (PRWZ02).

3.4 Related work

Evaluation has become an indispensible part of natural language processing research. The strong emphasis on solid evaluation has been seen as a way to move the field forward by identifying the most promising techniques and approaches. Specifically for the field of summarization, evaluation has received even more attention since there is no immediately obvious gold-standard or approach one can apply. The difficulties in summarization evaluation and suggested remedies have consistently accompanied the progress in automatic summarization (RRS61; MNP97; JBME98; GKMC99; DDM00).

Here we will overview some of the most recent approaches to evaluation presented and employed in research since 2000. These include DUC's method based on elementary discourse units, relative utility, information nuggets used for evaluation of definitional question answering systems, and factoids.

DUC content selection evaluation

The need for a large-scale evaluation within the Document Understanding Conference has generated a lot of discussion of evaluation and has provided large amounts of data for studies on evaluation. In the wake of the first conference in 2001 there was an active discussion on what approach should be used for intrinsic evaluation.

The final scheme that was adopted involved the comparison of peer summaries to a single human-authored model. The human model was automatically broken down into *elementary discourse units* (EDUs), capturing the need for analysis on a level smaller than a sentence. For each EDU, the human evaluator had to decide on a 1 to 5 scale the degree to which the peer expresses its information. In addition, for sentences in the peer that did not express any model EDU, the evaluators assigned a score reflecting whether the sentence contained important information.

Different proposals were made on how to incorporate the model EDU judgments into a final score, and average model EDU per summary was eventually adopted as a metric that was used throughout DUC 2004. Two of the main drawbacks of the approach were the use of a single model and the granularity of EDUs. For example, the post evaluation analysis of McKeown *et al.* (MBE^+01) indicated that *the model* had larger impact on peer scores than *which summarizer* performed the task. One of the possible remedies they suggested was the use of multiple models. In addition, Marcu (Mar01) reported that some systems were overly penalized since they contained content ranked as highly relevant for the topic, but not included in the model summary, again pointing out a shortcoming of the use of a single model.

The second drawback of the evaluation was the granularity of automatically identified EDUs. Marcu (Mar00) in Chapter 9 reports that in his work on developing and evaluating a discourse-based summarizer he, compared automatically identified elementary units with manually identified (by him) atomic meaning units in the text and remarks that "Usually, the granularity of the trees that are built by the rhetorical parser is coarser than the granularity of those that are built manually." This fact was confirmed by the NIST evaluators who reported having trouble deciding when an EDU can be considered matching content in the peer and were also unsure how to use context in order to interpret the meaning of SCUs. An example of the problems that annotators faced can be seen by considering the split into EDUs by a model sentence (from model D04.M.200.A.B.html):

[President Bush,]1 [after a poor initial start, has intensified efforts to bring aid to those affected.]2

The sentence is split into two EDUs delineated by square brackets and the evaluators had to provide judgments on how much of their content is expressed. The first model EDU contains a single name, while the second contains information about the poor initial start in the aid campaign, and the fact that more effort is currently made to improve the situation.

In our subsequent work on pyramid evaluation we tried to address the problems above, and we are grateful to the DUC organizers and participants for giving us the opportunity to analyze some of the problems and look for solutions.

Relative Utility

Relative utility (RJB00; RT03) was one of the possible evaluation approaches listed in the "Evaluation Road Map for Summarization Research"[4], prepared in the beginning of

[4] `www-nlpir.nist.gov/projects/duc/papers/summarization.roadmap.doc`

the Document Understanding Conferences. In this method, all sentences in the input are ranked on a scale from 0 to 10 as to their suitability for inclusion in a summary. In addition, sentences that contain similar information are explicitly marked, so that in the evaluation metric one could penalize for redundancy and reward equally informationally equivalent sentences. The ranking of sentences from the entire input allows for a lot of flexibility, because summaries of any size or compression rate can be evaluated. At the same time, the method is applicable only to extractive systems that select sentences directly from the input and do not attempt any reformulation or regeneration of the original journalist-written sentence.

The relative utility approach is very similar in spirit to the evaluation used by Marcu (Mar00), Chapter 9), who asked multiple independent subjects to rank the importance of information units following older research strategies (Joh70; Gar82). The main difference is that earlier research directly concentrated on subsentential units rather than sentences.

The data gathered by the relative utility experiments conducted by Radev and his colleagues was a very useful complement for our study on the importance distribution of content units of different granularities presented in this chapter.

Information nuggets

Information nuggets have served for evaluation of question answering systems on non-factoid questions, which require a longer answer, very similar to summarization.

Information nuggets are identified by human annotators through the analysis of *all systems' responses* to the question, as well as the searches made by the person who designed the question.

"An information nugget is an atomic piece of information about the target that is interesting (in the assessor's opinion). An information nugget is atomic if the assessor can make a binary decision as to whether the nugget appears in a response. Once the nugget list was created, the assessors marked some nuggets as vital, meaning that this information must be returned for a response to be good. Non-vital nuggets act as don't care conditions in that the assessor believes the information in the nugget to be interesting enough that returning the information is acceptable in, but not necessary for a good answer." (Voo04)

In theory, the requirement that information nuggets be atomic distinguishes nuggets from our SCUs. SCUs are of different granularity—usually highly-weighted SCUs are characterized by shorter contributors and more "atomicity" than lower-weight SCUs. As a consequence, annotators can experience more difficulties during peer annotation, since they have to use their own judgment on whether to match an SCU or not. The information nuggets are also specially tailored to the contents of peer answers and are, at least in theory, meant to be atomic with respect to peers. But when we look at actual question answering evaluations, the identification of nuggets in the systems' answer allows for a lot of freedom and subjective interpretation from the annotator.

The classification of nuggets into vital and non-vital is subjective, and as we already discussed would differ among different humans. In the question-answering settings, it is not possible to assign an empirical weight to a nugget, depending on the number of answers that contain it, since the nuggets are derived mainly from systems answers rather than from answers that a human would produce.

An example of a question, nuggets and system responses judged to contain the nuggets is shown here. The question asks for information about the sinking of a Russian submarine. The nuggets for the question, with their classification as "vital" and "ok" are given first and then five system responses judged as expressing the nugget "Norway provides rescue assistance" are listed. The example provided here is rather typical—information nuggets are usually quite general and can be expressed in many different ways in specific sentences.

```
Qid 66.8: 'other' question for target Russian submarine Kursk sinks
66 vital   Norway provides rescue assistance.
66 vital   Britain provides rescue assistance.
66 vital   Norwegian divers confirm flooding entire sub.
66 vital   Russian officials at first opine sinking resulted
           from collision.
66 okay    Two US subs monitored exercise involving Kursk.
66 okay    US sub in area of sinking heard explosion.
66 okay    Russians accept likelihood of on-board explosion.
66 okay    US provided their info to Russia re sinking.
```

```
66 okay    Russians critical of their officialdom's response.



-- The Norwegian ship Normand Pioneer with the British mini -
   submarine LR5 on board has left the operation area heading for
   Norway.

-- Russian navy said Saturday that all 118 members of the crew of the
   wrecked nuclear submarine Kursk are now likely dead, or will be
   before a British rescue submarine arrives with a team of Norwegian
   divers to make a final desperate effort to open a rear hatch and
   look for survivors.

-- A Norwegian admiral was quoted in a newspaper Thursday as saying
   that he had nearly called off Norway's efforts to rescue the crew
   of the Russian submarine Kursk because of Russian officials'
   interference and misinformation.

-- the joint british and norwegian began on sunday attempt to rescue
   any survivors on board the sunken russian submarine kursk, which is
   lying on the sea bed in the barents sea.

-- After reaching a Russian nuclear submarine that sank in the Barents
   Sea with 118 men on board, Norwegian divers struggled Sunday to
   open an escape hatch but found no signs of life.
```

It will be interesting to further explore the parallels of the pyramid method and the nugget-based evaluation approach, possibly combining desirable characteristics from both in order to reach a unified evaluation framework for non-factoid questions answering and summarization, as has already been suggested, for example, in Lin and Demner-Fushman (LDF05).

Factoid analysis

The most thorough analysis on the consensus of human summaries of the same text was presented by van Haltren and Teufel (vHT03). They collected 50 abstractive summaries of the same text and developed an annotation scheme for content units called *factoids*, analyzing the 50 abstracts in terms of factoids. Their large pool of summaries allowed for insightful observations and an empirical confirmation that the appearance of new content with the addition of new summaries does not tail off.

Their initial work was very semantically oriented, also including an analysis between the relations among different factoids, much in the spirit of the van Dijk tradition—"factoids correspond to expressions in a FOPL-style semantics, which are compositionally interpreted". They envisioned even further formalization of their mark-up—"First of all, the notation of the factoid (currently flat atoms) need to be made more expressive, e.g. by the addition of variables for discourse referents and events, which will make factoids more similar to FOPL expressions, and/or by the use of a typing mechanism to indicate the various forms of inference/implication". In their later work (TvH04b), where they included the analysis of another set of 20 summaries, they seem to settle to a representation closer to SCUs than on a first order logic language.

The two authors investigated in depth the annotation reliability for factoids and report remarkably good results in (vHT03; TvH04a; TvH04b) with kappa for factoid definition of 0.70 and even higher for factoid annotation—0.86 and 0.87. The annotations were done by the two authors themselves, and show that annotations of content can be done much more reliably by well-trained annotators.[5]

In their work, van Halteren and Teufel also address the question of *How many summaries are enough"* for stable evaluation results. Their investigation leads to the conclusion that 20 to 30 model summaries are necessary (in their EMNLP 2004 paper) and in the range of at least 30–40 (in their 2003 paper). This conclusion is dramatically different from the conclusion that we reached in our study of pyramid evaluation where we established that about 5 human models are necessary for stable results. A careful analysis shows that there is

[5] In this chapter we discussed the SCU annotation, quite similar to factoid annotation, by novice annotators in the context of DUC 2005 and the kappa for the novice annotators was 0.57.

no contradiction as it might seem at a first glance and that actually two different questions were addressed in their work and ours.

The approach that Teufel and Halteren take is the following: they resample their pool of summaries (with possible repetitions) in order to get sets of N summaries for different values of N.[6] Then for each pool of summaries derived in this manner, they score summaries against the factoid inventory using the weight of factoids in the peer summaries (without the normalization factor we propose). Then, for each pair of *system rankings*, regardless of the difference in scores, they compute the Spearman correlation coefficient and then take the average of the correlation coefficients for a given N. They deem a scoring reliable when the average correlation for a given N exceeds 0.95.

Our approach is more practically oriented—we assumed that a small difference in pyramid score does not necessarily entail a difference in summary quality. In fact, summaries with pyramid scores that differed by less than 0.06 were considered *equal* with respect to their information content (because such differences could be observed between annotations of the same summary by two different annotators). Then we proceeded to investigate what errors can arise in identifying summaries as being informationally equal or different. Consider for example the following scores for six systems under two different inventories of content units.

system	**Inventory 1**	**Inventory 2**
sys1	0.69	0.64
sys2	0.68	0.65
sys3	0.67	0.66
sys4	0.66	0.67
sys5	0.65	0.68
sys6	0.64	0.69

[6] They call this approach *bootstrap*, but it is not immediately obvious how the approach relates to the statistical method. One possible reason for concern is the fact that by resampling with repetition, they reach a situation in which a new summary does not add *any* new factoids to the factoid inventory. According to their own analysis, such an event is virtually impossible according to the empirical observations in their data.

In the pyramid analysis, all systems' summaries will be considered informationally equal under both inventories and thus the scores will be considered stable. But the rank correlation is perfectly negative, -1! So the apparent difference in conclusion in fact is due to the required strength of the results: in the Teufel and van Halteren study such a situation would be deemed as highly undesirable and unstable and in our study that had practical constraints in mind, the results will be good. But the fact that their stringent requirements were met when 30 or so human models were used indicated that a content unit based approach to summarization evaluation can have properties even beyond the practical evaluation requirements.

3.5 Conclusions

In this chapter we introduced the pyramid annotation procedure, which has two big advantages:

1. It allows to compare similarities and differences across multiple abstractive summaries for the same input. With this, it allows for empirical studies of human agreement on content selection.

2. It serves as a basis of an empirically motivated evaluation method that incorporates indications from multiple models to assign importance weight to different content.

The pyramid evaluation method leads to scores that are not influenced by the choice of model as long as enough models are used, as we showed in section 3.1. The scores are also diagnostic and allow system developers to see which was the good content in the input that their system did not include in the summary and change their system accordingly. As a result of the annotation, the summarization community is provided with a useful training corpus. The Ottawa University team produced such a corpus by tracing sentences in the evaluated machine summaries to the original document. Each sentence is annotated with a weight showing the relative importance of the sentence as a function of the content units it expresses. Such a training corpus is more useful than a binary annotation for each sentence on whether it should be included in a summary or not, because it reflects better that the content of an "ideal" summary is not unique.

The pyramid method for evaluation was applied on a large scale for the 2005 edition of the Document Understanding Conference. There, interannotator agreement could be computed in realistic settings, with annotations done by people who had not been previously exposed to the method. The overall difference in scores was in the expected range of 0.06, and the overall kappa statistic was 0.57. When content units of weight 1 were excluded from the calculations, the kappa statistic goes up to 0.65, very close the to generally recommended threshold of 0.67 at which the annotation can be considered reliable.

In the future, a group of experts will work on detailing and finalizing the guidelines for pyramid evaluation, which will lead to even broader acceptance of the method in the summarization community.

We analyzed the correlation between different pyramid scores and other evaluation metrics. A very high correlation between the original and modified pyramid scores, and the reduced annotation requirements for the modified scores, led to the recommendation that the modified score be used. In comparison with automatic metrics, there was not high enough correlation, so it does not yet seem feasible to replace the manual evaluation in multi-document summarization by the much cheaper and faster automatic method.

Finally, another direction for future work concerning evaluation is to study further human performance and develop tests that will distinguish good human summarizers from not so good or plain bad ones. In our work in this section, we assumed that all available models were created by good, equally competent summarizers. If we knew that one of the summarizers is better than the others, than content in his/her summaries will be more valuable and will have more weight that content coming from not so able summarizers. The DUC data is not very suitable for a study of this kind, because in it each summarizer wrote summaries for only a subset of all the data. Still, the DUC data can serve as a starting point for a study of differences in human summarization competance.

Chapter 4

Content selection: frequency, context and rewrite of generic noun phrases[1]

There has been growing interest in summarization in the past years, and in particular, the large amount of on-line news and information has led to the development of numerous multi-document summarizers for newswire[2], as well as online systems such as NewsInEssence (RBGZSR01)and the Columbia Newsblaster (MBE$^+$02) that run on a daily basis. The main problem an extractive summarization system needs to solve is content selection (i.e., deciding which sentences from the input documents are important enough to be included in a summary). Even systems that go beyond sentence extraction and use generation techniques to reformulate or simplify the text of the original articles need to decide which simplified sentences should be chosen, or which sentences should be fused together or rewritten (BME99; JM00; VBM04; BM05). The usual approach for identifying sentences for inclusion in the summary is to train a binary classifier (KPC95), a Markov model (CSGO04), or directly assign weight to sentences based on a number of features and heuristically chosen feature

[1] I started working on the material presented in this chapter during a summer internship at Microsft Research, with Lucy Vanderwende as my mentor. Our initial wok was described in (NV05)

[2] See for example `http://duc.nist.gov`

weights and pick the most highly weighted sentences (SNM02; LH02). But the question of which components and features of automatic summarizers contribute most to their performance has largely remained unanswered (MG01). In this chapter, we will examine several system-building decisions and the impact they have on the performance of generic multi-document summarizers of news. More specifically, we will research the following issues:

Frequency as indicator of importance In chapters 2 and 3 we examined multiple human written summaries to study the consensus between different writers on content selection decisions. Little predictable consensus could be seen, but observed human agreement was used to assign importance to different content units: the larger the number of people who express a content unit, the more important the content unit. The approach led to a good evaluation procedure that incorporates an empirical notion of importance. But an automatic summarizer does not normally have access to human summaries, and needs to estimate the importance of content from the input it receives, which is simply a set of news articles on the same topic. The question we pose here is "Does frequency in the input indicate importance?" Again, we use human summaries to define the gold-standard importance, and we show that content words (nouns, verbs and adjectives) and semantic content units that appear frequently in the input tend to also appear in human summaries and are thus indeed important. The detailed study of frequency as indicator of importance is presented in section 4.1.

Choice of composition function The frequency, and thus the importance, of content words can easily be estimated from the input to a summarizer. But is this enough to build a summarization system? Normally, a summarizer produces readable text as a summary, not a list of keywords, and needs to estimate the importance of larger text units, typically sentences. So a composition function needs to be chosen, that will estimate the importance of a sentence as a function of the importance of the content words that appear in the sentence. There are many possibilities for the choice of composition function, and in section 4.2 we will discuss three of them, showing that the choice can have a significant impact on the performance of the summarizer, ranging from close to baseline performance to state-of-the-art performance.

Context adjustment The notion of importance is not static: it depends on what has been already said in a summary. Take for example the DUC set that we show in figure 4.2, on the arrest of Augusto Pinochet in London on a Spanish extradition warrant. Initially, the most important information is the fact that Pinochet was arrested. After this fact has been included in the summary, other facts become *the currently most important*, for example that the arrest was in London or that it was on a Spanish arrest warrant. Context adjustment is especially important for multi-document summarization, where the input consists of many articles on the same topic. Several articles might contain sentences expressing the same information. It is possible that they all get high importance weights and if the summarizer does not have a module for context adjustment, or a duplication removal module at least, the summary would contain repetitive information. In section 4.2 we show how context adjustment improves content selection and reduces repetition in the summary.

Generic noun phrase rewrite Choosing a good composition function to combine the importance of content words into a measure of sentence importance leads to a summarizer with top performance. But the sentence is not necessarily the ideal granularity on which to judge importance. In section 4.3 we will discuss how the same ideas incorporated in the development of a summarizer that extracts sentences from the input can be used to develop a more flexible summarizer that alters the original author wording. We will show how computing importance of coreferential maximum noun phrases and choosing the best alternative leads to even further improvements in content selection.

We now proceed to a detailed discussion of these four aspects in the following sections: 4.1 (frequency in the input as an indicator for importance), 4.2 (choice of composition function and context adjustment), and 4.3 (rewrite of generic noun phrases). Frequency has been used as a feature in many summarization systems, but no study has isolated its impact on the system's performance and our task in these sections is to fill in this gap. We then present an overview of related work in development of multi-document summarizers in section 4.5 and close the chapter with a discussion of our findings in section 4.6.

4.1 Frequency and human summarization behavior

One of the issues studied since the inception of automatic summarization in the 60s is that of human agreement (RRS61): different people choose different content for their summaries (vHT03; RTSL03; NP04). More recently, others have studied the degree of overlap between input documents and human summaries (CS04; BV04). The natural question that arises if we combine the two types of studies is whether features in the input can allow us to predict what content humans would choose in a summary, and what content they would agree on. If such predictors are identified, they could be used as features for content selection by an automatic system. In this section, we focus on frequency, investigating the association between content that appears frequently in the input, and the likelihood that it will be selected by a human summarizer for inclusion in a summary. This question is especially important for the multi-document summarization task, where the input consists of several articles on the same topic and usually contains a considerable amount of repetition of the same facts across documents. We first discuss the link between frequency in the input at the word level and the appearance of words in human summaries (section 4.1.1), and then look at frequency at a semantic level, using semantic content units (section 4.1.2).

4.1.1 Content word frequency as indicator of importance

In order to study how frequency influences human summarization choices, we used the 30 test sets for the multi-document summarization task from DUC 2003. For each set, the input for summarization was available, along with four human abstracts for the input and the summaries produced by automatic summarizers that participated in the conference that year. Each of the inputs contained around 10 documents and the summaries were 100 words long. The counts for frequency in the input were taken over the concatenation of the documents in the input set.

Words frequent in the input appear in human summaries

We first turn to the question *Are content words that are very frequent in the input likely to appear in at least one of the human summaries?* We use a stop word list to exclude pronouns, function words and auxiliary verbs from consideration. Table 4.1 shows the percentage of the N most frequent content words from the input documents that also appear in the human models, for $N = 5, 8, 12$. In order to compare how many of these matches are achieved by a good automatic summarizer, we picked one of the top performing summarizers and computed how many of the N most frequent words from the input documents appeared in its automatic summaries, and the numbers are shown in the second row of table 4.1. For example, the table shows that, across the 30 sets, 95% of the five most frequent content words in the input were also used in at least one of the summaries, while the automatic summarizer used only 84% (first column of table 4.1).

	5 most frequent	**8 most frequent**	**12 most frequent**
used by human	94.66%	91.25%	85.25%
used by machine	84.00%	77.87%	66.08%

Table 4.1: Percentage of the N most frequent words from the input documents that appear in the four human models and in a state-of-the-art automatic summarizer (average across 30 input sets).

Two observations can be made about the table:

1. The high frequency words from the input are very likely to appear in the human models: the more frequent a word is in the input, the more likely it is that it will appear a human summary. This confirms that frequency is one of the factors that impacts a human's decision to include specific content in a summary. Using frequency in the input as indicator of importance probably helps the writers to resolve other different constraints such as personal interests and previous knowledge.

2. For the automatic summarizer, the trend to include more frequent words is preserved: the automatic summaries include 84% of the five most frequent words in the input, 78% of the 8 most frequent words, and 66% of the 12 most frequent. But the numbers are lower than those for the human summaries and the overlap between the machine summary and the human models can be improved if the inclusion of these most frequent words is targeted. As we will show later, it is possible to develop a summarizer

that includes a percentage of most frequent words equivalent to that in *four* human summaries. Trying to maximize the number of matches with the human models is reasonable, since on average across the 30 sets, the machine summary contained 30 content words that did not match any word in a human model.[3]

Humans agree on words that are frequent in the input

In the previous section we observed that the high frequency words in the input will tend to appear in *some* human model. But will high frequency words be words that the humans will agree on, and that will appear in *many* human summaries? In other words, we want to partition the words in the input into five classes C_n depending on how many human summaries they appear in, $n = 0...4$, and check if high class number is associated with higher frequency in the input for the words in the class. A word falls in C_0 if it does not appear in any of the human summaries, in C_1 if it appears in only one human summary and so on. Now we are interested to see how frequent the words in each class were in the respective input.

We found that, in fact, the words that human summarizers agreed to use in their summaries include the high frequency ones and the words that appear in only one human summary tend to be low frequency words as can be seen in table 4.2. The content words that were used by all four summarizers (in class C_4) had average frequency in the input equal to 31, while the words that never appeared in a human summary appeared on average about two times in the entire input of ten articles.

In the 30 sets of DUC 2003 data, the state-of-the-art machine summary contained 69% of the words appearing in all 4 human models and 46% of the words that appeared in 3 models. This indicates that high-frequency words, which human summarizers will tend to select and thus will be rewarded for example during automatic evaluation, are missing from the summary.

[3] Even though no rigorous study of the issue has been done, it can be considered that the content words that do not match any of the models describe “off-topic” events. This is consistent with the results from the quality evaluation of machine summaries in which human judges perceived more than half of the summary content to be “unnecessary, distracting or confusing.”

Class C_i	**Average $\lvert C_i \rvert$**	**Average frequency**
C_4	7	31
C_3	11	14
C_2	24	9
C_1	82	5
C_0	1115	2

Table 4.2: C_i (the first column) is the class of words that appear in i human summaries, Average $|C_i|$ (the second column) is the average size of class C_i, and the third column gives the average frequency of words in each class. The averages are computed for the 30 DUC'03 test sets.

Formalizing frequency: the multinomial model

The findings from the previous sections suggest that frequency in the inputs is strongly indicative of whether a word will be used in a human summary. We start out with assessing the plausibility of a formal method capturing the relation between the occurrence of content words in the input and in summaries by modeling the appearance of words in the summary under a multinomial distribution estimated from the input. That is, for each word w in the input vocabulary, we associate a probability $p(w)$ for it to be emitted into a summary. It is obvious that words with high frequency in the input will be assigned high emission probabilities.

The likelihood of a summary then is

$$L[sum; p(w_i)] = \frac{N!}{n_1!...n_r!} p(w_1)^{n_1} \cdot ... \cdot p(w_r)^{n_r} \tag{4.1}$$

where N is the number of words in the summary, $n_1 + ... + n_r = N$ and for each i, n_i is the number of times word w_i appears in the summary and $p(w_i)$ is the probability of w_i appearing in the summary estimated from the input documents. In order to confirm the hypothesis that human summaries have high likelihood under a multinomial model, we computed the log-likelihood $log(L[sum; p(w_i)])$ of all human and machine summaries from DUC'03 (see Table 4.3). The log-likelihood is computed rather than the likelihood in order to avoid numeric problems such as underflow for very small probabilities. If human summaries have higher likelihood under the model than machine ones, we can conclude that a multinomial model captures more aspects of the human summarization process than automatic summaries do. And indeed: the log-likelihood of summaries produced by human summarizers were overall higher than for those produced by systems and the fact that the top five highest log-likelihood scores belong to humans indicate that some humans indeed employ a summarization strategy informed by frequency.[4]

4.1.2 Frequency of semantic content units

We established that high-frequency content words in the input will be very likely to be used in human summaries, and that there will be a consensus about their inclusion in a summary between different human summarizers. But the co-occurrence of *words* in the inputs and the human summaries does not necessarily entail that the same *facts* have been covered. A better granularity for such investigation is the semantic content unit, an atomic fact expressed in a text, such as the summary content units we introduced in chapter 3.

Evans and McKeown (EM05) annotated 11 sets of input documents and human written summaries for SCUs following the pyramid approach. Based on their annotation, we were able to measure how predictive the frequency of content units in the documents is for the selection of the content unit in a human summary. As in our study for words, we looked at the N most frequent content units in the inputs and calculated the percentage of these that appeared in any of the human summaries. Similarly to the case of words, of the 5 most frequent content units, 96% appeared in a human summary across the 11 sets. The respective percentages for the top 8 and top 12 content units were 92% and 85%. Thus content unit frequency is highly predictive for inclusion in a human summary.

In addition, we computed the correlation between the weight of an SCU from the input documents (equal to the number of times the content unit was expressed in the input) and the SCU weight from human summaries (equal to the number of summarizers that expressed

[4] Other humans might have other strategies, such as giving maximum coverage of topics mentioned in the input, even such mentioned only once. Human10 appears to have such a strategy for example (after examination of his summaries).

Summarizer	**Log-likelihood**
Human1:	-198.65
Human2:	-205.90
Human3:	-205.91
Human4:	-206.21
Human5:	-206.37
System1:	-208.21
Human6:	-208.23
Human7:	-208.90
System2:	-210.14
System3:	-211.06
Human8:	-211.95
System4:	-212.57
System5:	-213.08
System6:	-213.65
Human9:	-215.65
System7:	-215.92
System8:	-216.04
System9:	-216.20
System10:	-216.24
System11:	-218.53
System12:	-219.21
System13:	-220.31
System14:	-220.93
System15:	-223.03
System16:	-225.20
Human10:	-227.17

Table 4.3: Average log-likelihood for the summaries of human and automatic summarizers in DUC'03. All summaries were truncated to 80 words to neutralize the effect of deviations from the required length of 100 words

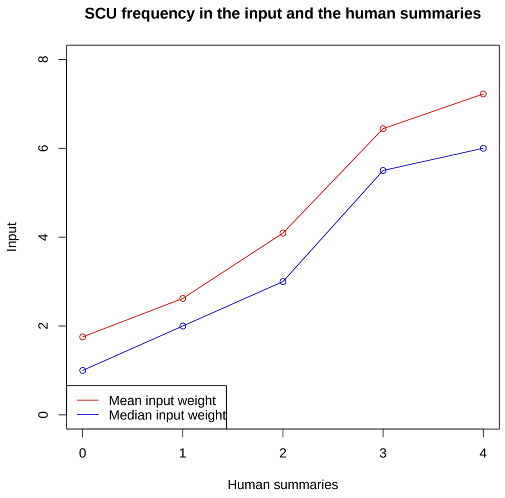

Figure 4.1: Input frequency of SCUs expressed in N human summaries. For $N = 0$ the SCU was not expressed by any of the summarizers.

the content unit in their summaries of the input. The Pearson's correlation coefficient between the input and human summaries weights is 0.64 (p=0), strongly indicating that content units that are repeated in several documents are likely to be picked in consensus by several humans. When we discussed mutually substitutable evaluation methods, we expected to see almost perfect correlation of 0.95 or higher. The observed correlation here is lower than perfect, but still shows that frequency in the input helps predict human agreement in terms of content units. The lower than perfect correlation shows that there are other factors at play that influence human content selection decisions, which we do not find surprising at all.

Figure 4.1 shows the median and mean frequency of an SCU in the input of about 10 articles, for the content units chosen from 0, 1, 2, 3, or 4 human summarizers (that is, it gives a plot of SCU frequency in the input versus SCU frequency in the four human summaries). There were 347, 95, 44, 34 and 27 content units in each respective class. The picture we see is exactly analogous to the one we saw in Table 3.2 when we investigated the question in terms of content word frequency. Content units that are expressed in more human summaries, also occurred more often in the input, in agreement with the conclusion we drew from the analogous investigation on the word level.

4.2 Composition functions and context adjustment

Now that we have convinced ourselves that frequency in the input is a good predictor of whether content will appear in human summaries and that human summaries have higher likelihood under a multinomial model, how can we extend these empirical findings to building a summarizer? The question is not trivial: normally, only the frequency of content words can be easily obtained from the input, but how is the frequency of words to be combined in order to get an estimate for the importance of sentences, the usual units for extraction in summarization?

We can define a family of summarizers, SUM_F, where F is the combination function that will give the importance of a sentence based on the words contained in that sentence. Different choices of F will give different summarizers from the frequency based summarizer family. Below we outline the overall summarization algorithm proposed in this dissertation and after that we will discuss possible choices of F.

Context-sensitive frequency-based summarizer

Step 1 Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input. The input has been parsed with Charniak's parser (Cha00) and only verbs, nouns, adjectives and numbers are considered in the computation of the probability distribution.

Step 2 Assign an importance weight to each sentence S_j in the input as a function of the importance of its content words.

$Weight(S_j) = F[p(w_i)]$ for $w_i \in S_j$

Step 3 Pick the best scoring sentence under the scoring function F from the previous step.

Step 4 For each word w_i in the sentence chosen at step 3, update its probability by setting it to a very small number close to 0.

Step 5 If the desired summary length has not been reached, go back to Step 2.

Different summarizers SUM_F can be obtained by making different choices for the composition function F. Three obvious candidates for F are:

Product ($F \equiv \prod$) For this choice of F, $Weight(S_j) = \prod_{w_i \in S_j} p(w_i)$

Average ($F \equiv Avr$) For this choice of F, $Weight(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{|\{w_i | w_i \in S_j\}|}$

Sum ($F \equiv \sum$) For this choice of F, $Weight(S_j) = \sum_{w_i \in S_j} p(w_i)$

Each of these choices for F leads to a different frequency based summarizer and we will see that the specific choice will have a huge impact on the performance of the summarizer, showing that not all frequency-based summarizers perform well.

Step 4 of the algorithm used in this thesis, the context adjustment step, also deserves some discussion. It serves a threefold purpose:

1. It gives the summarizer sensitivity to context. The notion of what is most important to include in the summary changes depending on what information has already been included in the summary.

2. By updating the probabilities in this intuitive way, we also allow words with initially low probability to have higher impact on the choice of subsequent sentences. If we look back at table 4.2, we see that this is a reasonable goal, since the large class of words that were expressed only in one model were not that frequent, so content that humans will not necessarily agree on, but is still good for inclusion, is not characterized by high frequency.

3. The update of word probability gives a natural way to deal with the redundancy in the multi-document input. No further checks for duplication seem to be necessary. In fact, in terms of content units, the inclusion of the same unit twice in the same summary is rather improbable.

Later in this chapter we will discuss what happens if *Step 4* is removed from the algorithm: it leads to worse content selection and significant increase in information repetition in the summary.

In *Step 1*, instead of choosing word classes based on their part-of-speech tag, we could use a simple stop word list in order to decide which words to be counted as content words. This version of SUM_F, with $F \equiv Avr$, has indeed been used in a summarization of machine-translated documents task, where parsing the input was not necessarily a reasonable approach. The light-weight version that uses a stop word list has the additional advantage of being faster. We will discuss the results of this evaluation later in this chapter.

It is of interest to see how a summarizer SUM_F does in terms of inclusion of top frequency words compared to humans and other top performing systems. Table 4.4 shows the percentage of the N most frequent words from the DUC'03 documents that also appear in SUM_{Avr} summaries. As expected, these are much higher than the percentages for the non-frequency oriented machine summarizer, but even higher than in all four human models taken together.

Summarizer	**5 most frequent**	**8 most frequent**	**12 most frequent**
human	94.66%	91.25%	85.25%
machine	84.00%	77.87%	66.08%
SUM_{Avr}	96.00%	95.00%	90.83%

Table 4.4: Percentage of the N most frequent words from the input documents that appear in one of the four human models, a state-of-the-art machine summarizer and SUM_{Avr}, a new machine summarizer based on frequency that uses the average as a composition function.

A summary produced by SUM_{Avr}, alongside the human summaries for the same set is shown in figures 4.2 and 4.3. For these summaries, the 20 most frequent words in the input are put in square brackets, indexed with the rank of the word. As suggested by the numbers in Table 4.4, SUM_{Avr} "over-generates" frequent words—it tends to include more of them than the human summarizers do. Specifically for the example in figure 4.2, there are only 5 of the top 20 words that SUM_{Avr} does not incorporate in the summary, while the four human summaries do not use 9, 9, 7, and 8 of the top rank words respectively. The figure also shows how the human summaries are very rich in high frequency words.

SUM$_{Avr}$ SUMMARY

While the [British]$_{10}$ [government]$_{7}$ defended the arrest, it and the [Spanish]$_{4}$ [government]$_{7}$ took no stand on [extradition]$_{9}$ of [Pinochet]$_{1}$ to [Spain,]$_{11}$ leaving it to the courts. A [Chilean]$_{3}$ specialist in international law was traveling to [London]$_{5}$ for further meetings with [British]$_{10}$ officials, Artaza [said.]$_{2}$ [Britain]$_{15}$ has defended its arrest of Gen. Augusto [Pinochet,]$_{1}$ with one lawmaker saying that [Chile's]$_{8}$ claim that the [former]$_{12}$ [Chilean]$_{3}$ [dictator]$_{14}$ has [diplomatic]$_{13}$ [immunity]$_{16}$ is ridiculous. He was arrested Oct. 16 at the instigation of a [Spanish]$_{4}$ magistrate seeking to extradite him on charges of genocide, terrorism and torture.

HUMAN SUMMARY 1

On Oct. 16, 1998 [British]$_{10}$ police arrested [former]$_{12}$ [Chilean]$_{3}$ [dictator]$_{14}$ [Pinochet]$_{1}$ on a [Spanish]$_{4}$ warrant charging murder of Spaniards in [Chile,]$_{8}$ 1973-1983. Fidel Castro denounced the [arrest.]$_{6}$ The [Chilean]$_{3}$ [government]$_{7}$ protested strongly. While the [British]$_{10}$ [government]$_{7}$ defended the [arrest,]$_{6}$ it and the [Spanish]$_{4}$ [government]$_{7}$ took no stand on [extradition]$_{9}$ of [Pinochet]$_{1}$ to [Spain,]$_{11}$ leaving it to the courts. [Chilean]$_{3}$ legislators lobbied in Madrid against [extradition,]$_{9}$ while others endorsed it. Then new charges were filed for crimes against Swiss and French citizens. [Pinochet's]$_{1}$ wife and family pleaded that he was too sick to face [extradition.]$_{9}$ As of Oct. 28 the matter was not resolved.

HUMAN SUMMARY 2

[Pinochet]$_{1}$ arrested in [London]$_{5}$ on Oct. 16 at a [Spanish]$_{4}$ judge's [request]$_{19}$ for atrocities against Spaniards in [Chile]$_{8}$ during his rule. Castro, [Chilean]$_{3}$ legislators and [Pinochet's]$_{1}$ lawyers protested and claimed he had [diplomatic]$_{13}$ [immunity.]$_{16}$ His wife asked for his release because he was recovering from recent back surgery. [Pinochet]$_{1}$ visited Thatcher before his surgery. The [British]$_{10}$ and [Spanish]$_{4}$ [governments]$_{7}$ defended the [arrest,]$_{6}$ saying it was strictly a legal matter. The EC president hoped [Pinochet]$_{1}$ would stand trial. None of his Swiss accounts have been frozen yet. The Swiss [government]$_{7}$ also asked for his arrest for the 1977 disappearance of a Swiss-[Chilean]$_{3}$ student.

Figure 4.2: Summaries produced for the same input by SUM$_{Avr}$ and by four human summarizers. The top twenty most frequent words are bracketed and the index shows their rank based on frequency in the input.

Continued:

HUMAN SUMMARY 3

[Former]$_{12}$ [Chilean]$_{3}$ [dictator]$_{14}$ Augusto [Pinochet]$_{1}$ has been arrested in [London]$_{5}$ at the [request]$_{19}$ of the [Spanish]$_{4}$ [government.]$_{7}$ [Pinochet,]$_{1}$ in London]$_{5}$ for back surgery , was arrested in his hospital room. [Spain]$_{11}$ is seeking [extradition]$_{9}$ of [Pinochet]$_{1}$ from [London]$_{5}$ to [Spain]$_{11}$ to face charges of murder in the deaths of [Spanish]$_{4}$ citizens in [Chile]$_{8}$ under [Pinochet's]1 rule in the 1970s and 80s. The arrest raised confusion in the international community as the legality of the move is debated. [Pinochet]$_{1}$ supporters say that [Pinochet's]$_{1}$ [arrest]$_{6}$ is illegal, claiming he has [diplomatic]$_{13}$ [immunity.]$_{16}$ The final outcome of the [extradition]$_{9}$ [request]$_{19}$ lies with the [Spanish]$_{4}$ courts.

HUMAN SUMMARY 4

[Britain]$_{15}$ caused international controversy and [Chilean]$_{3}$ turmoil by arresting [former]$_{12}$ [Chilean]$_{3}$ [dictator]$_{14}$ [Pinochet]$_{1}$ in [London]$_{5}$ for[Spain's]$_{11}$ investigation of [Spanish]$_{4}$ citizen deaths under [Pinochet's]$_{1}$ 17-year rule of torture and political murder. Claims are [Pinochet]$_{1}$ had [diplomatic]$_{13}$ [immunity,]$_{16}$ [extradition]$_{9}$ is international meddling or illegal because [Pinochet]$_{1}$ is not a [Spanish]$_{4}$ citizen, also his crimes should be punished. [Spain]$_{11}$ and [Britain,]$_{15}$ big [Chilean]$_{3}$ investors, fear damage to economic relations and let courts decide [extradition.]$_{9}$ The Swiss haven't investigated [Pinochet]1 accounts despite a [Spanish]$_{4}$ [request.]$_{19}$ [Pinochet]$_{1}$ is shielded from details, [said]$_{2}$ too sick to be extradited.

Figure 4.3: Summaries produced for the same input by SUM$_{Avr}$ and by four human. Only two of the top twenty words in the input do not appear in any of the summaries—[minister]$_{17}$ and [general]$_{18}$.

4.2.1 Evaluation results

To evaluate the performance of the three SUM_F summarizers, we use the test data from two large common data set evaluation initiatives—the 50 test sets for multi-document summarization task for DUC 2004 and the common test set provided in the 2005 Machine Translation and Summarization Evaluation initiative.

Document Understanding Conference

While we used the data from the 2003 DUC conference for development, the data from the DUC in 2004 was used as test data, which we report on here. We tested the SUM_F family of summarizers on the 50 sets from the generic summary task in 2004 DUC.

Even before proceeding to the usual summarization measure, we could see that the choice of combination function F has a significant impact on the summarizer performance. One would expect that the probabilistic summarizer $\text{SUM}_{\prod}$ would tend to favor shorter sentences because as the sentence gets longer, its overall probability involves the multiplication of more word probabilities (numbers between 0 and 1) and thus overall longer sentences will have lower probability. Exactly the opposite would be expected from $\text{SUM}_{\sum}$, which assigns sentences a weight equal to the sum of probabilities of the words in the sentence. The more words there are in the sentence, the higher the sentence weight will tend to be. SUM_{Avr} is a compromise between the two extremes. To confirm this intuition about the behavior of the summarizers depending on the choice of F, we looked at the length in sentences of the summaries that they produced. Table 4.2.1 shows the number of sentences across the 50 summaries produced by each of the systems. Our intuition is confirmed, with SUM_{Avr} producing summaries of about two sentences and $\text{SUM}_{\prod}$ getting about 6 sentences per summary, for the same size in words.

For the evaluation, we initially use the ROUGE-1 automatic metric, which has been shown to correlate well with human judgments based on comparison with a single model (LH03; Lin04) and which was found to have one of the best correlations with human judgment on the DUC 2004 data (OY04) among the several possible automatic metrics. In addition, we report the ROUGE-2 and ROUGE-SU4 metrics, which were used as official automatic evaluation metrics for MSE 2005 and DUC 2005.

System	Number of sentences	Sentences per summary
SUM$_\prod$	279	5.58
SUM$_{Avr}$	224	4.48
SUM$_\sum$	118	2.36

Table 4.5: Number of sentences in systems' summaries: the choice of composition function F affects systems' preference to longer or shorter sentences and SUM$_{Avr}$ is the more balanced one.

The results are obtained with ROUGE-1.5.5 with the settings used for DUC 2005 (with -*s* option for removing stopwords for ROUGE-1):

```
ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d
```

All summaries were truncated to 100 words (space delimited tokens) for the evaluation, as is normally done in DUC evaluations. The first column of table 4.6 also list the number of words in the 50 summaries in the test set. As can be noticed, some systems did not generate the longest possible summary. Peer 120 was an extreme example, producing summaries with average length of 78 words. But the impact of peer summary length on the final ranking of the systems is unlikely to be big, since most systems produced summaries very close to the required 100 word limit.

An approximate result on determining which differences in scores are significant can be obtained on the basis of comparing the 95% confidence intervals for each mean. Significant differences are those where the confidence intervals for the estimates of the means for the two systems either do not overlap at all, or where the two intervals overlap but neither contains the best estimate for the mean of the other (SG01).

Table 4.6 also shows scores for the 16 other participating systems from DUC 2004, and the baseline, which was selecting the beginning of the latest article as a summary.

Several conclusions can be drawn from the table:

Comparison between SUM$_F$ summarizers All three SUM$_F$ summarizers use word frequency in the input as a feature but have a different composition function F to assign weights to sentences. SUM$_\prod$ is a probabilistic summarizer and the weight it assigns

SYSTEM	ROUGE-1	ROUGE-2	ROUGE-SU4
SUM$_{\prod}$ (4900)	**0.305 (0.281; 0.329)**	**0.122 (0.108; 0.137)**	**0.159 (0.143; 0.175)**
peer 65 (4988)	0.305 (0.289; 0.320)	0.089 (0.081; 0.098)	0.130 (0.123; 0.137)
SUM$_{Avr}$ (4900)	**0.301 (0.277; 0.324)**	**0.110 (0.097;0.124)**	**0.146 (0.133; 0.158)**
SUM$_{AvrNoAdjust}$ (4900)	**0.296 (0.275; 0.319)**	**0.121 (0.105; 0.138)**	**0.159 (0.143; 0.175)**
peer 102 (4951)	0.285 (0.268; 0.303)	0.084 (0.076; 0.091)	0.126 (0.119; 0.132)
peer 34 (4954)	0.287 (0.271; 0.305)	0.074 (0.065; 0.083)	0.121 (0.113; 0.129)
peer 124 (4988)	0.282 (0.265; 0.300)	0.081 (0.073; 0.088)	0.123 (0.116; 0.131)
peer 44 (4854)	0.273 (0.256; 0.290)	0.076 (0.067; 0.084)	0.119 (0.111; 0.126)
peer 81 (4994)	0.268 (0.251; 0.285)	0.078 (0.070; 0.087)	0.121 (0.113; 0.128)
peer 55 (4971)	0.262 (0.247; 0.280)	0.069 (0.062; 0.077)	0.114 (0.107; 0.121)
peer 93 (4612)	0.253 (0.235; 0.271)	0.072 (0.066; 0.080)	0.107 (0.101; 0.114)
peer 120 (3903)	0.251 (0.231; 0.271)	0.077 (0.068; 0.085)	0.108 (0.099; 0.117)
peer 117 (4997)	0.238 (0.221; 0.257)	0.057 (0.051; 0.063)	0.107 (0.100; 0.113)
peer 140 (5000)	0.239 (0.219; 0.260)	0.068 (0.060; 0.076)	0.108 (0.101; 0.116)
peer 11 (4172)	0.239 (0.218; 0.259)	0.071 (0.062; 0.080)	0.105 (0.096; 0.114)
SUM$_{\sum}$ (4900)	**0.237 (0.217; 0.257)**	**0.070 (0.061; 0.081)**	**0.111 (0.101; 0.121)**
peer 138 (5000)	0.230 (0.211; 0.253)	0.069 (0.061; 0.077)	0.106 (0.098; 0.113)
Baseline (4899)	*0.202 (0.183; 0.221)*	*0.061 (0.052; 0.070)*	*0.098 (0.092; 0.106)*
peer27 (4686)	0.185 (0.166; 0.204)	0.046 (0.039; 0.055)	0.090 (0.083; 0.098)
peer123 (4338)	0.189 (0.173; 0.206)	0.049 (0.043; 0.056)	0.090 (0.084; 0.096)
peer 111 (5000)	0.063 (0.053; 0.073)	0.016 (0.013; 0.019)	0.057 (0.053; 0.061)

Table 4.6: DUC'04 ROUGE-1, ROUGE-2 and ROUGE-SU4 stemmed, stop-words removed for ROUGE-1 test set scores and their 95% confidence intervals for participating systems, the baseline, and SUM$_F$.

to each sentence is in fact the probability of the sentence. SUM_{Avr} and $\mathrm{SUM}_{\sum}$ assign to sentences weight equal to the average and the sum of the probabilities of the words in the sentence respectively. For these two latter summarizers, the raw frequency of words could be used instead of word probabilities. For all three automatic metrics, $\mathrm{SUM}_{\sum}$ is *significantly worse* than $\mathrm{SUM}_{\prod}$ and SUM_{Avr} and is in fact very close to baseline performance. Between SUM_{Avr} and $\mathrm{SUM}_{\prod}$, the probabilistic one ($\mathrm{SUM}_{\prod}$) is overall the better one and even on the ROUGE-SU4 metric it comes out *significantly* better than SUM_{Avr}.

Removing context adjustment In the fourth line in the table we have listed the automatic scores for $\mathrm{SUM}_{AvrNoAdjust}$. This is the summarizer for which the composition function $F \equiv Avr$, but without *Step 4* from the summarization algorithm responsible for adjusting the weights of words that appear in sentences already chosen for inclusion in the summary. The three automatic metrics give contradictory results about how the content selection capability of the summarizer is affected by the removal of the context adjustment step. According to ROUGE-1, removing the context adjustment leads to slightly lower results than if the adjustment is performed. For ROUGE-2, excluding the adjustment leads to slightly better results, and for ROUGE-SU4, excluding the adjustment step leads to *significantly* better results. We thus for the moment abstain from making a conclusion about the usefulness of content adjustment, and will address the question later in manual evaluation.

Comparison with other DUC systems $\mathrm{SUM}_{\prod}$ and SUM_{Avr} perform extremely well compared to the other DUC 2004 systems. In fact, according to ROUGE-2 and ROUGE-SU4, both of them are significantly better than *all* 16 other systems. The results from the ROUGE-1 metric are still good, but more modest: $\mathrm{SUM}_{\prod}$ is better than all but peer 65 and SUM_{Avr} has performance that is not significantly different from that of four of the other peers (peers 65, 102, 34, 124).

In sum, based on automatic scores (ROUGE-2 and ROUGE-SU4) we could claim that SUM_F summarizers outperform state-of-the-art systems. But using ROUGE-1, the strongest conclusion we can make is that SUM_F are about as good as the best systems. In order to

SYSTEM	Original	Modified	Repetition
peer 65	0.4435	0.3675	1.20
$\text{SUM}_{\prod}$	0.4063	0.3392	1.16
SUM_{Avr}	0.4039	0.3364	0.84
$\text{SUM}_{AvrNoAdjust}$	0.3811	0.3212	2.12
$\text{SUM}_{\sum}$	0.2614	0.1862	0.24

Table 4.7: Pyramid score results for peer 65 and SUM_F for the 50 DUC 2004 test sets. The original and modified pyramid scores are given in columns two and three, and the average number of repeated content units per summary is given in the fourth column.

more convincingly answer the question of where our summarizer stands with respect to the best DUC systems, we performed manual evaluation, using the pyramid method that we presented in chapter 3. All 50 test sets were annotated for SCUs and the summaries produced by the SUM_F summarizers and the best peer (peer 65), where evaluated manually. The results are shown in table 4.2.1. The original and the modified scores for each summarizer are listed, as well as the number of repeated content units in the summary in the fourth column labeled "Repetition".

The average score is highest for the best system at DUC'04, peer 65, but the difference between it and $\text{SUM}_{\prod}$ and SUM_{Avr} is not significant. $\text{SUM}_{\sum}$ is significantly worse than any of the other systems. Also, for the manual evaluation, the results we get for $\text{SUM}_{AvrNoAdjust}$ are similar to what we expected: without context adjustment, the performance of the summarizer gets somewhat worse, dropping from 0.4039 to 0.3811. But more importantly, with no context adjustment, the repetition of content units in the summary *increases significantly*, from 0.84 repeated SCU per summary for the summarizer with context adjustment to 2.12 repeated SCUs per summary for the same summarizer without the context adjustment step. The results on repetition also show that the probabilistic summarizer $\text{SUM}_{\prod}$ tends to include more repetition than SUM_{Avr}.

Overall, SUM_{Avr} is the best of the SUM_F family and this is the summarizer we choose for later comparisons. Its sentence selection scores comparable to that of the best DUC 2004 summarizer, it has most success in avoiding repetition in the summary from the frequency summarizer family, and is least sensitive to the influence of sentence length on the sentence weight.

It is interesting to note that from the automatic metrics, ROUGE-1, with stop words removed, leads to conclusions closest to those based on the manual evaluation. Since SUM$_F$ summarizers are based on word frequency, one could have expected the summaries would "game" the ROUGE-1 metric that computes word overlap between the models and the summary, but will do poorly on the other metrics. The results we see are totally controry to such expectations.

For the two pyramid scores, original and modified, we see that system rankings are the same, as we could expect since in chapter 3 we saw that the two metrics are mutually substitutable.

Machine translation and summarization evaluation 2005

In April 2005, a multi-document summarization evaluation task was conducted as part of the Machine Translation and Summarization Workshop at ACL.[5] The task was to produce a 100-word summary from multi-document inputs consisting of a mixture of English documents and machine translations of Arabic documents on the same topic. Some summarizers were especially modified to use redundancy to correct errors in the machine translations, or to avoid MT text altogether and choose only sentences from the English input. We ran SUM$_{Avr}$ without any modifications to account for the non-standard input (VS05). The light-weight version of the summarizer was run, which did not require part of speech tags and which excluded stop words from a given stop word list.

The official evaluation metrics adopted for the workshop were the manual pyramid score, ROUGE-2 (the bigram overlap metric) and R-SU4 (skip bigram). The skip bigram metric measures the occurrence of a pair of words in their original sentence order, permitting up to four intervening words. The metric was originally proposed for machine translation evaluation and was shown to correlate well with human judgments both for machine translation and for summarization (Lin04; LO04).

The pyramid method was used to evaluate only 10 of the test sets, while the automatic

[5] `http://www.isi.edu/~cyl/MTSE2005/MLSummEval.html`

metrics were applied to all 25 test sets. The average results for each peer for the three metrics is shown in table 4.8. For the manual pyramid scores, none of the differences between systems were significant according to a paired t-test at the 95% level of significance. This is not surprising, given the small number of test points. There were only three peers with average scores larger than that of SUM_{Avr}, and six systems with lower average pyramid performance. For the automatic metrics, significance was based again on the 95% confidence interval provided by ROUGE. One system was significantly better than SUM_{Avr}, and for each of the automatic metrics there were two systems that were significantly worse than SUM_{Avr}. The rest of the differences were not significant. In table 4.8, results that are significantly different from those for SUM_{Avr} are flagged by "***".

During the annotation for the pyramid scoring, the content units that were repeated in an automatic summary were marked up: we include in the results table the average number of repeated SCUs per summary for all systems. SUM_{Avr} was one of the systems with the lowest amount of repetition in its summaries, with three of the other peers including significantly more repetitive information. These results confirm our intuition that the weight update of words to adjust for context is sufficient for dealing with duplication removal problems. This experiment also confirms that SUM_{Avr} is a robust summarizer with good performance.

An interesting note can be made about the MSE results: peer 28 from this evaluation was the same system as peer 65 from the DUC 2004 evaluation. We describe this system in the related work section of this chapetr.

4.3 Rewrite of generic noun phrases

In the previous section we demonstrated that a frequency-based summarizer with appropriately chosen composition function and adjustment for context can achieve state-of-the-art performance in content selection for multi-document generic summarization. One of the drawbacks that we observed was that estimates for the importance of larger text units such as sentences depend on the length of the sentence. The natural question arises of whether the same approach of estimating importance can be applied to units smaller than sentences.

system	**pyramid**	**R-2**	**R-SU4**	**repetition**
1	0.52859	0.13076	0.15670	1.4
28	0.48926	0.16036***	0.18627***	3.4***
19	0.45852	0.11849	0.14971***	1.3
SUM$_{Avr}$	*0.45274*	*0.12678*	*0.15938*	*0.6*
10	0.44254	0.13038	0.16568	1.2
16	0.45059	0.13355	0.16177	0.9
13	0.43429	0.08580***	0.11141***	0.4
25	0.39823	0.11678	0.15079	2.7***
4	0.37297	0.12010	0.15394	4.1***
7	0.37159	0.09654***	0.13593	0.4

Table 4.8: Results from the MSE evaluation. Pyramid scores and duplication is computed for 10 test sets, automatic scores for all 25 test sets. Numbers flagged by "***" are significantly different from the results form SUM$_{Avr}$. For repetition, higher numbers are worse, indicating that there was more repetition in the summary.

The option to operate on smaller units, which can be mixed and matched from the input to give novel combinations in the summary, offers several possible advantages that we discuss next.

Improve content Sometimes sentences in the input can contain both information that is very appropriate to include in a summary and information that should not appear in a summary. Being able to remove unnecessary parts of the sentences can free up space for better information. Similarly, a sentence might be good overall, but could be further improved if more information on an entity or event is added in. Overall, a summarizer that is able to manipulate input sentences on subsentential units would theoretically be better at content selection.

Improve readability In chapter 2 we discussed the linguistic quality evaluation of automatic summaries and reported that summarizers perform rather poorly on several readability aspects, including referential clarity. In more than half of the automatic summaries there were entities for which it was not clear what/who they were and how they were related to the story. The ability to add in descriptions to entities in the summaries could improve the referential clarity of summaries and can be achieved by text rewrite of subsentential units.

IP issues Another very practical reason to be interested in altering the original wording of sentences in summaries in a news browsing system involves intellectual property issues. Newspapers are not willing to allow verbatim usage of long passages of their articles on commercial websites. Being able to change the original wording can thus allow companies to include longer than one sentences summaries, which as we discussed in chapter 2 would increase user satisfaction.

For these reasons, we now turn to exploring how the frequency-based summarizer framework can accommodate *rewrite of generic noun phrases*. That is, for a sentence in a summary, we will automatically examine the noun phrases in them and decide if a different noun phrase is more informative and should be included in the sentence in place of the original. Consider the following example:

Sentence 1 *The arrest* caused an international controversy.

Sentence 2 *The arrest in London of former Chilean dictator Augusto Pinochet* caused an international controversy.

Now, consider the situation where we need to express in a summary that the arrest was controversial and this is the first sentence in the summary, and sentence 1 is available in the input ("The arrest caused an international controversy"), as well as an unrelated sentence such as "The arrest in London of former Chilean dictator Augusto Pinochet was widely discussed in the British press". NP rewrite can allow us to form the rewritten sentence 2, which would be a much more informative first sentence for the summary: "The arrest in London of former Chilean dictator Augusto Pinochet caused an international controversy". Similarly, if sentence 2 is available in the input and it is selected in the summary after a sentence that expresses the fact that the arrest took place, it will be more appropriate to rewrite sentence 2 into sentence 1 for inclusion in the summary.

This example shows the potential power of generic noun phrase rewrite for summarization. It also suggests that context will play a role in the rewrite process, since different noun phrase realizations will be most appropriate depending on what has been said in the summary upto the point at which rewrite takes place.

4.3.1 NP-rewrite enhanced frequency summarizer

We can extend the summarization algorithm presented in the previous section to include NP rewrite. The inclusion of the new step presupposes that the classes of coreferring noun phrases in the input have been identified.

Step 1 Estimate the importance of content words based on their frequency in the input, $p(w_i) = \frac{n_i}{N}$.

Step 2 For each sentence S_j in the input, estimate its importance based on the words in the sentence $w_i \in S_j$.

Step 3 Select the sentence with the highest weight.

Step 4 For each maximum noun phrase NP_k in the selected sentence

4.1 For each coreferring noun phrase NP_i, such that $NP_i \equiv NP_k$ from all input documents, compute a weight $Weight(NP_i) = F_{RW}(w_r \in NP_i)$.

4.2 Select the noun phrase with the highest weight and insert in in the sentence. In case of ties, select the shorter noun phrase.

Step 5 For each content word in the rewritten sentence, update its weight by setting it to a number close to 0.

Step 6 If the desired summary length has not been reached, go to step 2.

Step 4 is the NP rewriting step. The function F_{RW} is the rewrite composition function that assigns weights to noun phrases based on the importance of words that appear in the noun phrase. The two options that we will explore here are $F_{RW} \equiv Avr$ and $F_{RW} \equiv \sum$. The two selections lead to different behavior in rewrite. $F_{RW} \equiv Avr$ will generally prefer the usage of short noun phrases, typically consisting of just the noun phrase head and it will overall tend to reduce the selected sentence. $F_{RW} \equiv \sum$ will behave quite differently: it will tend to insert information (add a longer noun phrase) when it has not been yet expressed in the summary, and will reduce the NP if the words in it already appear in the summary. This means that $F_{RW} \equiv \sum$ on the noun phrase level will have the behavior closest to what we expect for rewrite.

Of course, for an actual implementation of the algorithm, one needs a way to **identify maximum noun phrases** and to **identify coreference classes in the input**.

Maximum noun phrases

In dependency grammar formalisms (Mel88), the syntactic structure of a sentence is described in terms of a tree of words and syntactic relations between these words. The main verb of the sentence is normally the root of the dependency tree, and typical relations include **subj** (syntactic subject), **obj** (direct object), **dat** (indirect object), **attr** (premodifying nominals), **mod** (nominal postmodifiers) (TJ97). For example, the dependency representation of the sentence "British police arrested former Chilean dictator Augusto Pinochet" is given in figure 4.4.

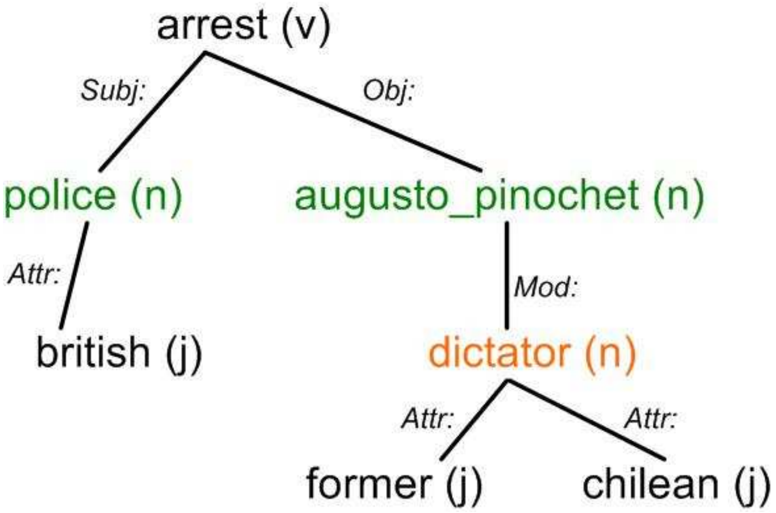

Figure 4.4: A dependency tree parse.

It is easy to define maximum noun phrases in a dependency representation: it is given by the subtree that has as a root a noun such that there is no other noun on the path between it and the root of the tree. For our example, there are two maximum NPs, with heads "police" and "Augusto_Pinochet"; "former chilean dictator" is not a maximum NP, since there is a noun (augusto_pinochet) on the path in the dependency tree between the noun "dictator" and the root of the tree.

The definition entails that a maximum NP includes all nominal and adjectival premodifiers of the head, as well as postmodifiers such as prepositional phrases, appositions, and relative clauses. This means that maximum NPs can be rather complex, covering a wide range of production rules in a context-free grammar. The dependency tree definition of maximum noun phrase makes it easy to see why these are a good unit for subsentential rewrite: the subtree that has the head of the NP as a root contains only modifiers of the head, and by rewriting the noun phrase, the amount of information expressed about the head entity can be varied.

In our actual implementation, a context free grammar probabilistic parser (Cha00) was used to parse the input. The maximum noun phrases were identified by finding a sequence of *<np>...</np>* tags in the sentence parse such that the number of opening and closing tags is equal. Each NP identified by such tag spans was considered as a candidate for rewrite.

Coreference classes

A coreference class CR_m is the class of all noun phrases in the text that refer to the same entity E_m. The general problem of coreference resolution is hard, and is even more complicated for the multi-document summarization case, in which cross-document resolution needs to be performed (Ng05; GA04). Since no freely available coreference resolution tool was available, we made a simplifying assumption, stating that all noun phrases that have the same noun as a head belong to the same coreference class. While we expected that this assumption would lead to some wrong decisions, we also suspected that in most common summarization scenarios, even if there are more than one entities expressed with the same noun, only one of them would be *main* for the news story and will be likely to be picked in a sentence for inclusion in the summary. We thus used the head noun equivalance to form the classes. A post-evaluation inspection of the summaries confirmed that our assumption was correct and there were only a small number of errors in the rewritten summaries that were due to coreference errors, which were greatly outnumbered by parsing errors for example. In a future evaluation, we will evaluate the rewrite module assuming perfect coreference and parsing, in order to see the impact of the core NP-rewrite approach itself.

4.4 NP rewrite evaluation

The NP-Rewrite summarization algorithm was applied to the 50 DUC 2004 test sets. Two examples of its operation with $F_{RW} \equiv Avr$ are shown below.

Original.1 While the British government defended *the arrest*, it took no stand on extradition of Pinochet to Spain.

NP-Rewite.1 While the British government defended *the arrest in London of former Chilean dictator Augusto Pinochet*, it took no stand on extradition of Pinochet to Spain.

Original.2 *Duisenberg* has said growth in the euro area countries next year will be about 2.5 percent, lower than *the 3 percent* predicted earlier.

NP-Rewrite.2 *Wim Duisenberg, the head of the new European Central Bank,* has said growth in the euro area will be about 2.5 percent, lower than *just 1 percent in the euro-zone unemployment* predicted earlier.

We can see that in both cases, the NP rewrite pasted into the sentence important additional information. But in the second example we also see an error that was caused by the simplifying assumption for the creation of the coreference classes according to which the percentage of unemployment and growth have been put in the same class. This example also suggests a possible way to improve the rewrite algorithm, even in the presence of coreference resolution errors. Mistakes can be avoided if rather than rewriting all noun phrases, rewrite were constrained to some subclass, excluding vague nouns such as *percent* in this example, and only rewriting nouns that are more specific.

In order to estimate how much the summary is changed because of the use of the NP rewrite, we computed the unigram overlap between the original extractive summary and the NP-rewrite summary. As expected, $F_{FW} \equiv \sum$ leads to bigger changes and on average the rewritten summaries contained only 54% of the unigrams from the extractive summaries; for $F_{RW} \equiv Avr$, there was much less change between the extractive and the rewritten summary, with 79% of the unigrams being the same between the two summaries.

Linguistic quality evaluation

As we mentioned in the beginning of this section, we hoped that the noun phrase rewrite will improve the referential clarity of summaries, by inserting in the sentences more information about entities when such is available. We were interested in how the rewrite version of the summarizer would compare to the extractive version, as well as how its linguistic quality compares to that of other summarizers that participated in DUC. Five summarizers were evaluated: peer 65, which is a (mostly) extractive summarizer that had the best content selection; peer 117, which was a system that used generation techniques to produce the summary and was the only real non-extractive summarizer participant at DUC 2004; the extractive frequency summarizer with average as a composition function for importance of sentences, and the two versions of the rewrite algorithm. The evaluated rewritten summaries had potential errors coming from different sources, such as coreference resolution, parsing errors, sentence splitting errors, as well as errors coming directly from rewrite, in which an unsuitable NP is chosen to be included in the summary. Improvements in parsing for example could lead to better overall rewrite results, but we evaluated the output as is, in order to see what is the performance that can be expected in realistic setting for fully automatic rewrite.

The evaluation was done by five native English speakers, using the five DUC linguistic quality questions on grammaticality, repetition, referential clarity, focus and coherence. Five evaluators were used so that possible idiosyncratic preference of a single evaluator could be avoided. Each evaluator ranked all five summaries for each set, presented in a random order. The results are shown in table 4.4. Each summary was evaluated for each of the properties on a scale from 1 to 5, with 5 being very good with respect to the quality and 1, very bad.

These results allow us to do the system comparisons we were interested in.

Comparing the extractive summarizers The first two rows give the scores for the

SYSTEM	**Grammar**	**Repetition**	**References**	**Focus**	**Coherence**
peer 65	4.30	3.98	4.22	3.80	3.38
SUM$_{Avr;Id}$	4.06	4.12	3.80	3.80	3.20
SUM$_{Avr;Avr}$	3.40	3.90	3.36	3.52	2.80
SUM$_{Avr;\sum}$	2.96	3.34	3.30	3.48	2.80
peer 117	2.06	3.08	2.42	3.12	2.10

Table 4.9: Linguistic quality evaluation results for five systems: peer 65 is the best performing system at DUC 2004, peer 117 was the only generative systems; SUM$_{Avr;Id}$ is the frequency summarizer with no NP rewrite; and the two versions of rewrite with sum and average as combination functions.

two extractive summarizers, peer 65 and SUM$_{Avr;Id}$, the frequency summarizer without rewrite. These two systems are expected to have good grammaticality scores, since they do not attempt any modification of the original sentences. Both systems have an average score of around 4 (good) for grammatically. There are several reasons why the systems are not closer to the best score of 5. First, the summaries were truncated to 100 words, so sometimes the last sentence in the summary was not complete. Second, there were occasional errors in sentence splitting, and when the systems choose a sentence that had been incorrectly split, they introduce ungrammatical fragments in the summary. Finally, on some occasions, the original journalist-written sentences were long and somewhat difficult to read and the evaluators ranked summaries that included such sentences lower on grammaticality, even though they were not ungrammatical per se, but simply employed cumbersome grammar. The two summarizers are overall indistinguishable from one another on linguistic quality, with the exception of the clarity of reference. The average score for reference clarity for peer 65 is higher than that of the frequency summarizer by 0.42, which is a significant difference. The results on repetition confirm the independent information collected by the pyramid evaluation, with the frequency summarizer doing a bit better in avoiding repetition than peer 65.

Comparing NP rewrite to extraction Here we would be interested in comparing the extractive frequency summarizer (SUM$_{Avr;Id}$), and the two version of systems that rewrite noun phrases: $\mathrm{SUM}_{Avr;Avr}$ (which changes about 20% of the text) and $\mathrm{SUM}_{Avr;\sum}$ (which changes about 50% of the text). The general trend that we see for all five dimensions of linguistic quality is that the more the text is automatically altered, the worse the linguistic quality of the summary gets. In particular, the grammaticality of the summaries drops significantly for the rewrite systems. The increase of repetition is also significant between $\mathrm{SUM}_{Avr;Id}$ and $\mathrm{SUM}_{Avr;\sum}$. Error analysis showed that sometimes increased repetition occurred in the process of rewrite for the following reason: the context weight update for words is done only after each noun phrase in the sentence has been rewritten. Occasionally, this led to a situation in which a noun phrase was augmented with information that was expressed later in the original sentence. The referential clarity of rewritten summaries also drops significantly, which is a rather disappointing result, since one of the motivations for doing noun phrase rewrite was the desire to improve referential clarity by adding information where such is necessary. One of the problems here is that it is almost impossible for human evaluators to abstract themselves from grammatical errors when judging referential clarity. Grammar errors decrease the overall readability of the summary and a summary that is given a lower grammar ranking tends to also receive lower referential clarity score, increasing the challenge for summarizers that move towards abstraction and alter the original wording of sentences.

Comparing $\mathrm{SUM}_{Avr;\sum}$ and peer 117 We finally turn to the comparison of between $\mathrm{SUM}_{Avr;\sum}$ and the generation based system 117. This system is unique among the DUC 2004 systems, and the only one that year that experimented with generation techniques for summarization. System 117 analizes the input in terms of predicate-argument triples and identifies the most important triples. These are then verbalized by a generation system originally developed as a realization component in a machine translation engine. Thus, peer 117 made even more changes to the original text then the NP-rewrite system. The results of the comparison are consistent with the observation that the more changes are made to the original sentences, the more the readability of summaries decreases. $\mathrm{SUM}_{Avr;\sum}$ is significantly better than peer 117 on all five readability aspects, with notable difference in the grammaticality and referential quality, for which $\mathrm{SUM}_{Avr;\sum}$ outperforms peer 117 by a full point. This indicates that NP rewrite is a good candidate granularity for sentence changes and it can lead to significant altering of the text while preserving significantly better overall readability.

General discussion of findings We saw that in terms of linguistic quality, extractive systems will be superior at the current point of research development to systems that alter the original wording from the input. Moreover, extractive and abstractive systems are evaluated together and compared against each other, putting pressure on system developers and preventing them from fully exploring the strengths of generation techniques. It seems that if researchers in the field are to explore non-extractive methods, they would need to compare their systems separately from extractive systems, at least in the beginning exploration stages. The development of non-extractive approaches in absolutely necessary if automatic summarization were to achieve levels of performance close to human, given the highly abstractive form of summaries written by people.

Another observation seen from the evaluation is that both extractive and non-extractive systems perform rather poorly in terms of the focus and coherence of the summaries that they produce. We did discuss this fact already in chapter 2, but the proposed frequency based method did not show any improvement over other extractive systems. One of the steps for future work would be to outline a proposal for integrating modules that would improve the focus and coherence of summaries, and to test it out.

Content selection evaluation

In the previous section we discussed the linguistic quality of extractive summaries and summaries produced using NP rewrite and full-scale generation. We now turn to examine the question of how the content in the summaries changed due to the NP-rewrite, since improving content selection was the other motivation for exploring rewrite. In particular, we are interested in the change in content selection between $\mathrm{SUM}_{Avr;\sum}$ and $\mathrm{SUM}_{Avr;Id}$ (the extractive summarizer). We use $\mathrm{SUM}_{Avr;\sum}$ for the comparison because it lead to bigger changes in the summary text compared to the purely extractive version. We used the pyramid evaluation method to address the question. Of the 50 test sets, there were 22 sets in which the NP-rewritten version had lower pyramid scores than the extractive version of the summary, 23 sets in which the rewritten summaries had better scores, and 5 sets in which the rewritten and extractive summaries had exactly the same scores. So we see that in half of the cases the NP-rewrite actually improved the content of the summary. The summarizer version that uses NP-rewrite has overall better content selection performance than the purely extractive system. The table below shows that the original pyramid score increased from 0.4039 to 0.4169 for the version with rewrite.

The table below shows the pyramid scores for $\text{SUM}_{Avr;Id}$ and $\text{SUM}_{Avr;\sum}$.

System	Original pyramid score	Modified pyramid score
$\text{SUM}_{Avr;\sum}$	0.4169	0.3420
$\text{SUM}_{Avr;Id}$	0.4039	0.3364

Here is an example of pairs of summaries for the same set in which the NP-rewritten version had better content. After each summary, we list the content units from the pyramid that were expressed in the summary. Before the label of each content unit, we list the weight, and content units that differ between the extractive and rewritten version are displayed in italic. The rewritten version conveys high weight content units that do not appear in the extractive version, with weights 4 and 3 respectively.

Fully extractive summary Italy's Communist Refounding Party rejected Prime Minister Prodi's proposed 1999 budget. By one vote, Premier Romano Prodi's center-left coalition lost a confidence vote in the Chamber of Deputies Friday, and he went to the presidential palace to rsign. Three days after the collapse of Premier Romano Prodi's center-left government, Italy's president began calling in political leaders Monday to try to reach a consensus on a new government. Prodi has said he would call a confidence vote if he lost the Communists' support." I have always acted with coherence," Prodi said before a morning meeting with President Oscar Luigi.

- (4) Prodi lost a confidence vote
- (4) The Refounding Party is Italy's Communist Party
- (4) The Refounding Party rejected the government's budget
- (3) The dispute is over the 1999 budget
- (2) Prodi's coalition was center-left coalition

- (2) The confidence vote was lost by only 1 vote
- (1) Prodi is the Italian Prime Minister
- *(1) Prodi wants a confidence vote from Parliament*

NP-rewrite version Communist Refounding, a fringe group of hard-line leftists who broke with the minstream Communists after they overhauled the party following the collapse of Communism in Eastern Europe rejected Prime Minister Prodi's proposed 1999 budget. By only one vote, the center-left prime minister of Italy, Romano Prodi, lost The vote in the lower chamber of Parliament 313 against the confidence motion brought by the government to 312 in favor in Parliament Friday and was toppled from power. President Oscar Luigi Scalfaro, who asked him to stay on as caretaker premier while the head of state decides whether to call elections.

- (4) Prodi lost a confidence vote
- *(4) Prodi will stay as caretaker until a new government is formed*
- (4) The Refounding Party is Italy's Communist Party
- (4) The Refounding Party rejected the government's budget
- *(3) Scalfaro must decide whether to hold new elections*
- (3) The dispute is over the 1999 budget
- (2) Prodi's coalition was center-left coalition
- (2) The confidence vote was lost by only 1 vote
- (1) Prodi is the Italian Prime Minister

Another example, which showed the worse deterioration of the rewritten summary compared to the extractive one, both in terms of grammaticality and content, is shown below. Here, the problem with repetition during rewrite arises: the same person is mentioned twice in the same sentence and at both places the same overly long description is included in the sentence, rendering it practically unreadable.

Fully extractive summary Police said Henderson and McKinney lured Shepard from the bar by saying they too were gay and one of their girlfriends said Shepard had embarrassed one of the men by making a pass at him. 1,000 people mourned Matthew Shepherd, the gay University of Wyoming student who was severely beaten and left to die tied to a fence. With passersby spontaneously joining the protest group, two women held another sign that read," No Hate Crimes in Wyoming." Two candlelight vigils were held Sunday night. Russell Anderson, 21, and Aaron McKinney, 21, were charged with attempted murder.

- (4) The victim was a student at the University of Wyoming
- *(4) The victim was brutally beaten*
- *(4) The victim was openly gay*
- *(3) The crime was widely denounced*
- (3) The nearly lifeless body was tied to a fence
- *(3) The victim died*
- *(3) The victim was left to die*
- *(2) The men were arrested on charges of kidnapping and attempted first degree murder*
- *(2) There were candlelight vigils in support for the victim*
- (1) Russell Henderson and Aaron McKinney are the names of the people responsible for the death

NP-rewrite version Police said Henderson and McKinney lured the The slight, soft-spoken 21-year-old Shepard, a freshman at the University of Wyoming, who became an overnight symbol of anti-gay violence after he was found dangling from the fence by a passerby from a bar by saying they too were gay and one of their girlfriends said the The slight, soft-spoken 21-year-old Shepard, a freshman at the University of Wyoming, who became an overnight symbol of anti-gay violence after he was found dangling from the fence by a passerby had embarrassed one of the new ads in that supposedly hate-free crusade.

- (4) The victim was a student at the University of Wyoming

- (3)The nearly lifeless body was tied to a fence
- *(1) A passerby found the victim*
- (1) Russell Henderson and Aaron McKinney are the names of the people responsible for the death
- *(1) The victim was 22-year old*

Even from this unsuccessful attempt for rewrite we can see how changes of the original text can be desirable, since some of the newly introduced information is in fact interesting for the summary.

4.5 Related work

We will now overview several other approaches for generic multi-document summarization.

Description of peer 65 In the previous sections we compared the performance of our summarizer to that of the best system in DUC 2004. We saw that SUM$_{Avr}$ achieves a comparable level of performance, both in content selction and linguistic quality. It is of course of interest to see what techniques are used in that summarizer. So before we turn to a general overview of related work, we will devote this section to a detailed description of peer 65 and a comparison between it and SUM$_F$.

Peer 65 has particpated in all DUC evaluations between 2001 and 2004 (CSOO01; SOC$^+$02; DCS$^+$03; CSGO04). And has evolved over the years from mediocre performance to the best system. The system is a supervised hidden Markov model that uses topic signature words as features. The HMM assigns scores to each sentence and then the best sentences that cover the largest number of signature words form the summary.

We now in turn discuss each component of the system and outline the similarities and differences between it and SUM$_F$.

Training an HMM Unlike our data-driven algorithm, peer 65 uses a supervised method, HMM, that needs to be trained on manually annotated data. The hidden stucture of the model allows the summarizer to factor in some context in the sentence ranking process: namely, it takes into account whether the sentence preceeding the currently scored sentence was classified as a summary sentence or not. The only other feature that the system uses is a function of the number of topic signature words in the sentence: $log(n+1)$, where n is the number of signature terms in the sentence. The HMM then assigns a posterior probability to each sentence for it to be included in the summary, and these posterior probabilities can be considered as sentence weights. Across the different years of DUC, modifications were made in order to find the best structure for the HMM. Also, in the beginning of DUC, the system also included a number of other features such as sentence position in the document, term frequency, and grammatical categories of words, but as these features were dropped, the performance of the summarizer improved, until eventually the frequency related likelihood ratios for the topic signature terms remained the only feature. We next explain the idea of topic signartures.

Topic signature terms The idea of topic signatures was introduced by Lin and Hovy (LH00) in the context of single document summarization, and was later used is several multi-document summarization systems, including in peer 65.

Lin and Hovy's idea was to automatically identify words that are descriptive for a cluster of documents on the same topic, such as the input to a multi-document summarizer. We will call this cluster T. Since the goal is to find descriptive terms, a collection of documents not on the topic is also necessary (we will call this background collection NT). To this end, one can use the likelihood ratio statistic introduced in the field of computational linguistics by (Dun94). The idea is to define a probabilistic model of the data, so that we can make statistical inference and decide which terms t are associated with T more strongly than with NT than one would expect by chance. This is done in the following way:

There are two possibilities for the distribution of a term t: either it is very indicative of the topic of cluster T, and appears more often in documents associated with T than in other documents NT, or the term t is not related to the topic and appears with equal frequncy across both T and NT. These two alternatives can be formally written as two hypothesis about the possible state of affairs:

H1: $P(t|T) = P(t|NT) = p$ (t is not a descriptive term for the topic)

H2: $P(t|T) = p_1$ and $P(t|NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

Now, let us look at the collection of the background documents and the topic cluster as a sequence of words w_i: $w_1 w_2 \dots w_N$. We are interested in occurances of t, that is, in the cases when some $w_i = t$. The occurance of each word is thus a Bernoulli trail with probability p of success, with success when $w_i = t$ and failure otherwise. Then the overall probability of observing the term t appearing k times in the N trails is given by the binomial distribution

$b(k, N, p) = \binom{N}{k} p^k (1-p)^{N-k}$

This model allows us to compute the likelihood of the data (the T+NT corpus) under both hypotheses. For that purpose, we need the actual estimates of p, p_1 and p_2. The maximum likelihood estimates are

$p = \frac{c_t}{N}$, where c_t is equal to the number of times term t appeared in the entire corpus T+NT, and N is the number of words in the entire corpus.

Similarly, $p_1 = \frac{c_T}{N_T}$, where c_T is the number of times term t occured in T and N_T is the number of all words in T.

$p_2 = \frac{c_{NT}}{N_{NT}}$, where c_{NT} is the number of times term t occurred in NT and N_{NT} is the total number of words in NT.

We can define $\lambda = \frac{\text{Likelihood of the data given H1}}{\text{Likelihood of the data given H2}}$

λ can be computed directly when we have all the necessary counts of occurance of t and the number of words in ech corpus.

Oftentimes, the quality $-2log\lambda$ is computed rather then λ since it has a well-know distribution: χ^2. Bigger values of $-2log\lambda$ indicate that the likelihood of the data under H2 is higher, and the χ^2 distribution can be used to determine when it is significantly higher (when $-2log\lambda$ exceeds 10).

So, for terms t for which the computed $-2log\lambda$ is higher than 10, we can infer that they occur more often with the topic T than in a general corpus NT, and we can dub them "topic signature terms".

Two final notes about $-2log\lambda$ are due:

1. This quantity is computed *as a function of the frequency of the term t in the topic documents*. It can be seen as an alternative weight for the term, different from the maximum likelihood that we used in SUM$_F$, but still a function of the term frequency in the input.

2. Its calculation requires an *additional corpus* of off-topic documents.

In peer 65, *an empirically set treshold* is used as a cut-off value and all terms in the input that have likelihood ratio that exceeds this treshold are dubbed signature terms. Then, for the HMM model, a function of the number of such terms is used as a feature and a posterior probability for inclusion in the summary is assigned to the sentence.

Pivoted QR decomposition This step incorporates the idea of context adjustment for peer 65. This is a formal algebraic model that allows the weights of sentences to be updated based on previous choices in the summary. A matrix A is formed, with columns corresponding to each sentence in the input, and rows corresponding to each word in the input, with entry a_{ij} equal to the weight of term i if it appears in sentence j and zero otherwise. The sentence with highest posterior probability is chosen for the summary. The norm of all other sentences (columns) is then reduced proportionally by substracting off the component of the column that lies in the direction of the column that was just added. Overall, the pivoted QR decompositions seems like a complex formal model of the context adjustment step that we proposed for the SUM$_F$ family. As we saw from the results on repetition on the DUC and MSE data, the step in SUM$_F$ seems to be more effective, resulting in less inclusion of repetitive information in the summary.

In summary, peer 65 is characterized by the following features:

1. It assigns weights on words as a complex function of the frequency of the word in the input.

2. It has two paramaters that are empirically set: the number of states in the HMM and the cut-off value of the likelihood ratio above which a term is considered to be a signature term.

3. It is a supervised algorithm, and requires manually annotated training data, in contrast to SUM$_F$ that is fully data driven from the input to the summarizer.

4. It requires an additional background corpus for the computation of the likelihood ratio for terms in the input. In DUC runs, the remaining test sets were used as such a corpus. It is not clear how and if the the selection of the background corpus will affect the final results outside of DUC runs (for example in the context of online news browsing system).

In comparison, SUM$_F$ achieves comparable results without any use of additional data besides the input (no background corpus and no training data), it does not have any manually set tresholds, and its conceptual simplicity allows for the better evaluation and understanding of how the different components contribute to the summarizer performance. One open question that could be addressed in future research is to compare the list of most frequent words in the set to the words with highest likelihood ratio. The complication of requiring a background corpus and of comuting the likelihood ration will be justified only if the two lists differ.

Lite_GISTexter Lite_GISTexter (LHHN04) was one of the well performing algorithms in DUC 2004 and was developed with the goal of producing a conceptually simple but reliable multidocument summarizer. It assigns weights for each term in the input, equal to $-2log\lambda$ as we defined it in the previous section in the discussion of topic signature words. Each sentence is assigned a weight equal to the *sum* of words that appear in the sentence. Duplication removal was addressed with the following heuristic: the highest scoring sentence was added to the summary, and then the next highest ranked sentence gets added to the summary if the number of topic signature terms that it has in common with sentences already found in the summary is less than the number of signature terms that it introduces in the summary. This heuristic aims at approximating the idea of Maximal Marginal Relevance (CG98) which prescribes that sentences that are relevant but maximally different from sentences already in the summary should be used to achieve wider coverage.

So Lite_GISTexter also falls in the class of frequency-based summarizers: it assigns weights to words (using a background corpus), then combines the weights of words into weights of sentences, and picks the top ranking sentences using a variant of MMR to reduce duplication. Interestingly, in their DUC report, the Lite_GISTexter team reports that they tried alternative weighting schemes that did not work as well as the likelihood ratio, concluding that frequency in the input is not a good feature. As we saw in our experiments, indeed sum is a very bad choice of composition function and leads to bad content selection results. But we also showed that this does not mean that frequency in the input is not a good feature, since with the better choice of composition function, the results for SUM$_F$ improve much more, with SUM$_\prod$ significantly outperfroming Lite_GISTexter on all three automatic evaluation metrics.

DEMS/NEATS/MEAD DEMS (SNM02) is a flexible multi-document summarizer that has been evaluated favorably in all DUC evaluations from 2001 to 2004, and that runs as part of the Newsblaster online news browsing and summarization system. DEMS uses several features to compute sentence weight, combining these features into a single weight using manually asigned linear combination coefficients. The features used in the system include the *location* of the sentence in the input (sentences that appear later in the documents are penalized), publication date (giving preference to later publications), *named entities*, *pronoun* (sentences that contain pronouns are penalized. It also uses frequency, but rather than computing frequency of individual, it computes the frequency of *concepts*, which is a collection of related terms. Other novel features used in DEMS are a special dictionary of words that were found to be charactersitic of opening paragraphs of news articles and of a dictionary of verb specificity that contains highly imformative verbs. Many of these features are intuitively very interesting, but unfortunately it is not clear how much each feature contributes to the good performance, and how much the different additional features improve over frequency alone. This way of arbitrary combining of weights for different features is also typical for other systems, for example NEATS (LH02), and MEAD (RBGZ01). The non-modular architecture of these systems make the flexibility of the SUM_F summarizers stand out. As we mentioned, we know that such systems use frequency as a feature, but we do not know how much of the summarizers' performance is explained by the frequency feature. These systems use similarity measures based on word overlap to remove redundancy in the summary: the similarity between a sentence and the already chosen parts of the summary is computed. If it exceeds some predetermined treshold, the sentence is rejected, and the treshold is manually assigned by the researchers.

Sentence clustering approaches One of the first, and very popular, approaches to multi-document summarization was to cluster topically related sentences from the input and select one sentence from the cluster as a representative of the topic in the summary (BME99; BKN01; SNM04). These summarizers obviously try to exploit frequency on the sentence level, with clusters with more sentences considered more important. Again, a hidden parameter can change the results considerably, since if lower similarity between sentences in the cluster is required, bigger clusters can be formed, but the sentences in them will not be tightly related on the same topic. Such an approach of assigning importance to sentences also deals directly with the problem of duplication removal: since only one sentence per cluster is chosen, the summary would not include repetition. Interestingly, the size of the cluster (equivalent to sentence frequency), did not lead to good content selection performance. The problem was addressed by adding in the weighting of clusters lexical chains (in (BME99)) and *tf.idf* (for (SNM04)). The addition of such information, which incorporates in the cluster score the frequency also of the words in the sentences, leads to much better results in content selection.

Graph-based algorithms Some of the most newly developed multi-document summarizers are those that reduce the problem of summarization to graph problems, notably using the PageRank algorithm (ER04b; MT04; VBM04). Of these, the most successful application to multi-document summarization was that of Erkan and Radev. In their LexRank algorithm, each sentence defines a node in the text graph. To define edges in the graph, the cosine similarity between two sentences is computed and an edge is added between the nodes representing the two sentences if the similarity exceeds a predetermined treshold. Thus the edges are defined for sentences that share the same words. The PageRank algorithm (PBMW98) is then used iteratively to compute the rank (importance) of each sentence as a function of the number of neighbors and the importance of the neighbors of each node. The iterations distribute the weight across the graph, and quickly converge to stable node weights. In their discussion of the approach, the authors explain that the iterative spreading of importance in the graph is similar to a voting process: sentences from the entire graph vote for sentences with which they share word overlap. Of course, such a voting procedure can be achieved by a direct frequency count, rather than distributing importance little by little through the nodes. So the PageRank algorithm can be seen as a complex (unobservable) function that assigns weights to sentences based on the frequency of words that appear in the text. In order to avoid repetition, sentences that are assigned high importance, but are similar to more important sentences are not included in the summary.

Approaches to abstractive summarization

The potential benefit of abstractive summarization has been long recognized in the summarization community. Jing and McKeown, for example, developed the cut-and-paste approach for single-document summarization, which relies on analysis of professional human abstracts to identify text transformation operations used by humans (JM00). They identified six operations that can be used alone or together to transform extracted sentences in human-written abstracts. Sample operations included sentence reduction, sentence combination, syntactic transformation, and lexical paraphrasing. The implementation of operations includes multiple knowledge sources: large lexicons with subcategorization information (to ensure grammaticality); context information (phrases that have links with surronding sentences are not removed because they are considered more important); corpus evidence (the probability that a human would remove or reduce a phrase). The approach worked very well for single-document summarization, but research has shown (BV04) that these operations are not enough to explain the much more complex text modifications that humans perform when summarizing multiple articles.

We next overview several approaches tried for abstractive summarization of multiple documents.

peer 117 In the previous section we compared the readability aspects of the summaries produced by a generative event-based summarizer (VBM04), peer 117. In their approach, they have as a goal to identify important *events* as expressed by important verbs, rather than enities and explore how focusing on events would change the nature of the obtained summaries. In order to approach the problem in an event-centric manner, they use a graph scoring algorithm to identify highly weighted nodes and relations in a graph constructed using a dependency-style analysis for the input documents. The scoring is used to guide content selection, which is then presented for realization to a generation component originally developed as a realization module in a machine translation system. Our apporach to NP-rewrite, in contrast, is entity-centered, modifying the amount of information expressed about a particular entity in the summary, rather than a particular event.

Information fusion This technique was developed to address in a specific way the problem of sentence redundancy in multi-document summarization (BM05). Given a set of similar sentences, information fusion produces a new sentence containing the information common to most sentences in the set, thus reducing sentence length and avoiding source bais. To identify common information, fusion uses a bottom-up local multisequence alignment of dependency trees, using words and paraphrases as anchors. Combination of fragments is addressed through construction of a fusion lattice encompassing the resulting alignment and linearization of the lattice into a sentence using a language model. So fusion is a sophisticated tool for text-to-text generation, but it does not use any ranking of the importance of the information in the sentences since its goal is different.

Sentence compression and simplification Sentence compression, a technique for removing certain fragments from a sentence without rendering it ungramatical has been a challanging research topic, even outside of summarization applications. Chandrasekar *et al.* (CDS96) viewed text simplification as a preprocessing tool to improve the performance of their parser. The PSET project (CMP$^+$99), on the other hand, focused its research on simplifying newspaper text for aphasics, who have trouble with long sentences and complicated grammatical constructs. Rule-based systems have been developed (Gra98; Sid02; Sid03), as well as statistical models trained on a corpus of human reductions (Jin00; KM00; RKCZ03), and discourse-informed methods have been proposed recently (SL05). But overall, little integration has been done for these approaches for direct use in summarization. The major challange to integrating compression methods in summarization systems is that it is difficult to incorporate considerations of importance in the simplification process, the way Jing did for single document summarization and we proposed here for NP-rewrite in multi-document summarization. Research has shown that simplification of background information in the form of relative clauses and appositions can be useful to improve content selction (SNM04; CSGO04), but the findings have not yet been confirmed in a complete manual evaluation.

The Trimmer system (ZDL$^+$05) has recently been developed, extending a system for headline generation into a system for topic-focused multi-document summarization. Again, for them the main challange is to decide how to incorporate importance of textual units into the trimming process. For example, they do not attempt to trim out clauses that contain named entities. Their results in content selection could be possibly improved if they used a frequency based estimation for importance.

4.6 Discussion and conclusions

In this chapter, we proposed and evaluated a family of context-sensitive, frequency based summarizers. The summarizers we described are unsupervised and data driven: the only feature they use is word frequency in the input. This feature has been traditionally used in summarization, but its relative impact on system performance had not been studied before. We motivated the use of the feature with analysis of human summaries, both on the word and content unit level, and demonstrated that indeed human summarizers tend to choose for their abstracts words and content units that appear frequently in the input, and that there is more agreement between humans for more frequent words and content units.

The proposed algorithm has simple well-defined steps, which allow us to examine how the choice of a composition function for assigning weights to sentences and how context adjustments are made changes the performance of the system. Different composition functions lead to differently performing summarizers, ranging from close to baseline to state-of-the-art performance. Our results showed that taking average for a composition function gives the most balanced summarizer, that does not choose too long or too short sentences and has very good performance on content selection and linguistic quality. We also saw that context adjustment is an obligatory step for a good summarizer, because it improves content selection and reduces the amount of repeated information in the summary.

SUM$_{Avr}$ outperforms most of the other 16 summarizers from the DUC 2004, and compares favorably with the best performing DUC 2004 system. Manual evaluation showed that SUM$_{Avr}$ has slightly lower performance than the best system (0.40 vs 0.44), but the difference is not statistically significant. SUM$_{Avr}$ significantly outperforms the best system in its ability to remove duplication from the summary. These results are very positive, because the state-of-the-art system is supervised, and needs manually annotated training data, as well as a background corpus used to compute term weights. It also uses two empirically set parameters (number of states for the HMM and topic signature word likelihood ratio cut-off). Finally, the theoretical complexity of the state-of-the-art system clouds the understanding of what really makes the system work. As we showed here in the related work section, it, as well as several other multi-document summarizer, also falls in the general class of frequency based summarizers.

Since the SUM$_F$ algorithm yielded good results for content selection, we explored an extension for noun phrase rewrite, again using word frequency to estimate the importance of maximum noun phrases and choose the best alternative for the context. NP-rewrite led to summaries that were between 20% and 50% different from their purely extractive versions. Our goal for exploring rewrite was to improve clarity of references in the summary and improve content selection through the additional flexibility. Unfortunately, our results showed that the linguistic quality of summaries drops with the use of NP-rewrite. Still, NP-rewrite significantly outperforms a fully generative event-centric system on grammaticality, referencial clarity and focus. In terms of content selection, NP-rewrite did improve content in 23 out of the 50 test sets, but the overall improvement across all sets was not significant, rsisng from 0.40 to 0.41. While the results from these experiments were not as good as we would have liked, showed two very important points. First, that the importance of smaller text units can be computed as a function of term frequency, thus suggesting a way to enhance existing text simplification approaches, for example. Second, our experiments show that when rewrite techniques are fully intergrated in a summarizer and evaluated against extractive summarizers, they will tend to have lower readability scores. Which means that in order to encourage the further development of abstractive summarization systems, it would be beneficial to compare text altering systems seperately from extractive systems, so that researchers can explore solutions for the widely recognized need for non-extractive approaches to summarization, rather than develop extractive variations that would stay competative against others.

In summary, we have proposed a conceptually simple algorithm for generic multi-document summarization that achieves state of the art perfoprmance. We hope it can be used as a banchmark for future developers of multi-document systems, so that they can explore the contributions of novel features, and linguistically motivated processing. We also hope to further explore the possibilities for rewrite.

Chapter 5

References to people in summarization

Apporaches to generic multi-document summarization need to be robust: the input to a summarizer can be a collection of articles on any topic, ranging from political events, through natural diseasers, to scientific break throughs. The question that we pose in this section is whether, in the settings of an unrestricted domain task, we can apply linguistically motivated analysis to enhance the summarization output. In this chapter, we demonstrate that this is indeed possible. We make use of the fact that new reports, and subsequently summaries, are centered around people. For example, about 30% of all searches in a web search engine contained a person's name, asking for information about a person (GG04). Also, in the DUC data, there were on average 3.85 references to people per 100-word human summary; hence it is important for news summarization systems to have a way of modeling the cognitive status of such referents and a theory for referring to people.

It is also important to note that there are differences in references to people between news reports and human summaries of news. Journalistic conventions for many mainstream newspapers dictate that initial mentions to people include a minimum description such as their role or title and affiliation. However, in human summaries, where there are greater space constraints, the nature of initial references changes. We observed (SNM04) that in DUC'04 and DUC'03 data[1], news reports contain on average one appositive phrase or relative clause every 3.9 sentences, while the human summaries contain only one per 8.9 sentences on average. In addition to this, we observe from the same data that the average length of a first reference to a named entity is 4.5 words in the news reports and only 3.6 words in human summaries. These statistics imply that human summarizers do compress references, and thus can save space in the summary for presenting information about the events. Cognitive status models can inform a system when such reference compression is appropriate.

So, for this chapter, we focus on references to people, the charactersitrics of the references and the appropriate form of reference. In human communication, the wording used by speakers to refer to a discourse entity depends on their *communicative goal* and their beliefs about *what listeners already know*. The speaker's goals and beliefs about the listener's knowledge are both a part of a cognitive/mental model of the discourse.

Cognitive status distinctions depend on two parameters related to the referent—*a)* whether it already exists in the hearer's model of the discourse, and *b)* its degree of salience. The influence of these distinctions on the form of referring expressions has been investigated in the past. For example, centering theory (GJW95) deals predominantly with local salience (local attentional status), and the givenness hierarchy (information status) of Prince (Pri92) focuses on how a referent got in the discourse model (e.g. through a direct mention in the current discourse, through previous knowledge, or through inference), leading to distinctions such as discourse-old, discourse-new, hearer-old, hearer-new, inferable and containing inferable. Gundel *at al.* (GHZ93) attempt to merge salience and givenness in a single hierarchy consisting of six distinctions in cognitive status (in focus, activated, familiar, uniquely identifiable, referential, type-identifiable).

Among the distinctions that have an impact on the form of references in a summary are the *familiarity* of the referent:

D. Discourse-old vs discourse-new

[1] The data provided under DUC for these years includes sets of about 10 news reports, 4 human summaries for each set, and the summaries by participating machine summarizers.

H. Hearer-old vs hearer-new

and its global salience:

M. Major vs minor

The notion of global salience is very important to summarization, both during content selection and during generation on initial references to entities. For example, in chapter 4 we showed that globally salient words and content units that are likely to appear in human summaries tend to appear frequently in the input and are thus part of the topic, or aboutness, of the news story. On the other hand, local salience (*in focus* or *local attentional state*) is relevant to anaphoric usage during subsequent mentions. While the global salience has been discussed in the linguistic literature, here we can take a definition more specific to the process of summarization through the use of human summaries. We define globally saleint entities from the input to be those that are also mentioned in a human summary. The globally salient entities are the *Major* characters of the news story.

In general, initial (discourse-new) references to entities are longer and more descriptive, while subsequent (discourse-old) references are shorter and have a purely referential function. In section 5.1, we show the results of a corpus study, that confirm that indeed discourse-new references are characterized by different syntactic form than discourse-old entities, we show how these differences in form can be formally modeled and how it can be used to automatically rewrite summaries to achieve better fluency and readability.

The other two cognitive status distinctions, whether an entity is central to the summary or not (major or minor) and whether the hearer can be assumed to be already familiar with the entity (hearer-old vs hearer-new status), are discussed in section 5.2. There is a tradeoff, particularly important for a short summary, between what the speaker wants to convey and how much the listener needs to know. The hearer-old/new distinction can be used to determine whether a description for a character is required from the listener's perspective. The major/minor distinction plays a role in defining the communicative goal, such as what the summary should be about and which characters are important enough to refer to by name.

5.1 Rewrite of references to people

Automatically generated summaries, and particularly multi-document summaries, suffer from lack of coherence (BN00). One explanation for this fact is that the most widespread summarization strategy is still sentence extraction, where sentences are extracted word-for-word from the original documents and are strung together to form a summary. While some researchers have developed methods to regenerate summary text from the text of the original articles (e.g., (BME99; Jin00; KM02; SMC01), the focus has been mostly on removing irrelevant and redundant phrases or on fusing information from different articles (BM05; EKKM05).

Outside of summarization, though, different aspects of coherence have been studied in great detail. In particular, seminal work on centering (GJW95) motivated numerous investigations of the factors that influence the *local coherence* of discourse. Centering theory looks at two main sources of (in)coherence—the syntactic realization of discourse entities and the transition between focused entities. These studies are pertinent for summarization: Barzilay *et al.* (BEM02), for example, have shown how considerations of the latter kind can be used to guide ordering in multi-document summaries. The work most directly related to summary rewrite is that of Mani *et al.* (MGB99), in which summaries are revised by aggregating together information on the same entity and extraneous descriptions are dropped. Radev and McKeown (RM97) also emphasized the importance of references to people in summaries, and described how information on people can be collected over the web and used as an additional information source during summarization. But syntactic form and its influence on summary readability have not been taken into account in the implementation of a full-fledged summarizer.

Considerations of local coherence are extremely important for summaries, which are very short by definition and thus, are less affected by deficiencies in global discourse structure.[2] Figure 5.1 shows a summary generated by the Columbia Summarizer (MBE$^+$02). The summary gives a good idea of what incoherence problems can arise—the first mention of

[2] But as discussed in Chapter 2, in longer summaries of 250 words, global aspects such as organization and focus become more problemtic.

the two politicians in the summary does not use the first mention of these entitites in any of the original articles and this makes it difficult for the reader to know to whom the summary refers.

> **Terrell** had 56 percent of the white vote to 31 percent for **Landrieu**, while Landrieu had 75 percent of the black vote to 10 percent for Terrell. A poll released this week shows the race between **Democratic Sen. Mary Landrieu** and **her Republican challenger, Suzanne Haik Terrell**, to be dead even. Voters go to the polls Saturday. With Louisiana's Senate run-off election just four days away, President Bush led the GOP charge Tuesday for **Republican candidate Suzanne Haik Terrell** in what polls now suggest is a toss-up race against **freshman Democratic Sen. Mary Landrieu**.

Figure 5.1: A problematic summary

These difficulties of text comprehension due to inappropriate syntactic forms have been discussed previously. One of the main claims of centering theory is that different syntactic realizations pose different processing requirements on the hearer and thus contribute to the coherence of discourse. Krahmer and Theune (KT02) report an experiment on human preference to sequences of syntactic forms that demonstrates that people prefer subsequent mentions that are less informative than the previous mentions of the same entity. They also cite experiments that show that utterances are more difficult to read if a definite description or a proper name is used in places where a pronoun is appropriate (GGG93).

Here we conduct a corpus study to identify the syntactic properties of first and subsequent mentions of people in newswire text. The resulting statistical model of the flow of referential expressions in text is based on features that can be derived from full text using shallow parsing technology. Thus, it can be used to create a set of recommended rewrite rules that can transform the summary back to a more coherent and readable text. Our study focuses on noun phrases containing mentions of people names. These constitute a subset of the general problem of reference in summaries that exemplify the general problem of under and overspecification in reference. Yet, restriction to people's names allows us to build a working solution due to recent advances in langugage technology, namely statistical parsing and named entity recognition.

In the following sections, we first describe the corpora that we used and then two statistical models that we developed for the task. The first is based on Markov chains and models how subsequent mentions are conditioned by earlier mentions, while the second, stratified model captures the different types of realizations for first through fifth mention separately. We close with discussion of our evaluation, which measures how well the models can regenerate the sequence of references in a test corpus, demonstrating that the Markov model is far more informative.

5.1.1 The Corpus

We used a corpus of news from the test data used in DUC 2001 through 2003, containing 651,000 words drawn from six different newswire agencies, in order to study the syntactic form of noun phrases in which references to people have been realized. The variety of sources was used, because working with text from one specific source could lead to the learning of the paper-specific editorial rules. We use the full input documents rather than summaries, because the documents contain many more reference to people than summaries, as we discuss in the next section.

We were interested in the occurrence of features such as type and number of premodifiers, presence and type of postmodifiers, and form of name reference for people. We began our study by manually annotating a small corpus of six articles; this pilot study allowed us to determine which feature of interest could be automatically extracted. We then constructed a large, automatically annotated corpus by merging the output of Charniak's statistical parser (Cha00) with that of the IBM named entity recognition and coreference system Nominator (WRC97). The automatically derived corpus contains references to 6240 distinct entities as recognized by the named entity coreference system.

In this section, we describe the features that were annotated.

Given that we restricted references to mentions of people, there are two distinct types of premodifiers, "titles" and "name-external modifiers". The titles are capitalized noun premodifiers that conventionally are recognized as part of the name, such as "president" in "President George W. Bush." (Cha01), for instance, discusses statistical techniques for disambiguating name structure in examples like the one above, and shows how the structure can be parsed to identify the first name, the last name, middle initial and title modifiers.

Name-external premodifiers are modifiers that do not constitute part of the name, such as “irish flutist” in “Irish flutist James Galway”.

The three major categories of postmodification that we distinguish are apposition, prepositional phrase modification and relative clause. All other postmodifications, such as remarks in parentheses and verb-initial modifications are lumped in a category “others”.

We identified four categories of names corresponding to the general European and American name structure. They include full name (first + last name), middle initial (first name + middle initial + last name), last name only, and nickname (first name or nickname).

Examples of the different properties coded for noun phrases are given in Figure 5.2. In sum, the features of the target NP that we examined were:

- Is the target named entity the head of the phrase or not?
- Is it in a possessive construction or not?
- If it is the head, what kind of pre- and post- modification does it have?
- How was the name itself realized in the noun phrase?

In order to identify the appropriate sequences of syntactic forms in coreferring noun phrases, we analyze the coreference chains for each entity mentioned in the text. A coreference chain consists of all the mentions of an entity within a document. In the manually built corpus, a coreference chain can include pronouns and common nouns that refer to the person. However, these forms could not be automatically identified, so coreference chains in the automatically derived corpus only include noun phrases that contain at least one word from the name. There were 3548 coreference chains in the automatically derived corpus; an example is given in Figure 5.3 which shows both the full coreference derived manually and the abbreviated chain identified in the automatically derived corpus.

5.1.2 Statistical Models of Mention Sequence

We developed two models of syntactic realization. The first uses a Markov chain model; it represents the influence of each mention on the subsequent reference. The second models the likelihood of particular forms of syntactic realization for first and subsequent mentions

NP1: John Aquilino, a former NRA official who now publishes his own gun owners' newsletter.

Codes as: full name + apposition

NP2: Chief Petty Officer Luis Diaz of the U.S. Coast Guard in Miami.

Coded as: 3 title premodifiers + full name + prepositional phrase postmodification

NP3: Dutch speed skater Yvonne van Gennip

Coded as: 3 name-external premods + middle initial name

NP4: Soviet pianist Vladimir Feltsman, who arrived in the United States last August after an eight-year battle to emigrate,

Coded as: 2 name-external + full name + relative clause

NP5: Powell, stationed behind the group of reporters who were questioning Reagan during an Oval Office photo opportunity,

Coded as: last name + other postmodification

Figure 5.2: Examples of different syntactic forms

separately. Our results show that the the Markov chain model is more informative than the stratified model.

The Markov Chain Model

The initial examination of the data showed that syntactic forms in coreference chains can be nicely modeled by Markov chains.

The formal definition of a Markov chain follows:

Let X_n be random variables taking values in I. We say that $(X_n)_{n\geq 0}$ is a Markov chain with initial distribution λ and transition matrix P if

- X_0 has distribution λ
- for $n \geq 0$, conditional on $X_n = i$, X_{n+1} has distribution $(p_{ij}|j \in I)$ and is independent of $X_0, ..., X_{n-1}$.

Informally, a Markov chain is given by a transition matrix and an initial distribution. The transition matrix gives the probability of moving from one state to the next, while

Bill Clinton X

the newly installed President

The man, whom almost two-thirds of all Americans trusted

Bill Clinton X

Clinton X

he (2)

Clinton X

he

Clinton (5) X

Bill Clinton X

Clinton (2) X

the president

Clinton (2) X

Bill Clinton X

the President

Figure 5.3: Full coreference chain for a person. The number in paretheses shows how many times the given syntactic form has been repeated consecutively in the chain. The entries shown in bold and marked with an "X" represent the chain that will be derived automatically.

the initial distribution gives the probability of being in a specific state at time zero. The probability of being in a given state at a given time depends only on the probability of the preceding state at the previous time. All these properties have very visible counterparts in the behavior of coreference chains. The first mention of an entity does have a very special status ((Fra90) and (PV98)) and its appropriate choice makes text more readable. Thus, the initial distribution of a Markov chain would correspond to the probability of choosing a specific syntactic realization for the first mention of a person in the text. For each subsequent mention, the model assumes that only the form of the immediately preceeding mention determines its form. This property of the model will be tested in evaluation, but seems to predict intuitively plausible sequences. For example, if a person has been previously mentioned by full name, then it is most likely appropriate to refer to him again by his last name, but if the previous mention was a last name, then a subsequent last name mention is appropropriate with even higher probability.

Of course, additional discourse factors can play a role in determining the use of a given type of NP. For example, Fox (Fox98) and Levy (Lev84) give detailed studies of the global context factors that play a role in syntactic realization. For example, in longer discourse, an entity that as already mentioned, but was not in focus in the preceeding discourse segment get re-introduced with descriptions that are different from the first mention in the discourse and is still richer than a subsequent mention that would occur within the same segment. But for now, we will adopt the simple Markov chain model to see how useful it can be.

	modification	no modification
initial	**0.76**	0.24
modification	0.44	0.56
no modification	0.24	**0.75**

Figure 5.4: Markov chain for modification transitions. The first row gives the initial distribution vector.

	full name	last name	nickname
initial	**0.97**	0.02	0.01
full name	0.20	**0.75**	0.05
last name	0.06	**0.91**	0.02
nickname	0.24	0.22	**0.53**

Figure 5.5: Markov chain for name realization. The first row gives the initial distribution vector.

The Stratified Model

The stratified model is guided by the idea that it is not just the first mention that has special characteristics, but rather that there are special features of the syntactic form strongly associated with the first mention, other features associated with the second mention and so

	apposition	none	prepositional	relcl	other
initial	**0.25**	0.60	*0.07*	*0.04*	*0.04*
apposition	0.06	0.88	*0.00*	*0.04*	*0.02*
none	0.04	0.89	*0.01*	*0.03*	*0.03*
prepositional	0.10	0.80	*0.01*	*0.07*	*0.02*
relcl	0.08	0.82	*0.01*	*0.06*	*0.03*
other	0.07	0.88	*0.00*	*0.04*	*0.01*

Figure 5.6: Markov chain for postmodification.

	0	1	2	3	4	5	6
initial	0.49	0.22	0.16	0.08	0.03	0.01	0.01
0	0.86	*0.09*	*0.04*	*0.01*	*0.00*	*0.00*	*0.00*
1	0.43	0.50	*0.05*	*0.01*	*0.00*	*0.01*	*0.00*
2	0.78	0.13	0.08	*0.01*	*0.01*	*0.00*	*0.00*
3	0.78	0.13	0.07	0.01	*0.01*	*0.00*	*0.00*
4	0.74	0.09	0.15	0.02	0.00	*0.00*	*0.00*
5	0.90	0.10	0.00	0.00	0.00	0.00	*0.00*
6	0.81	0.06	0.13	0.00	0.00	0.00	0.00

Figure 5.7: Markov chain for the number of premodifiers. Count given for merged title and external premodifiers.

on. Since summaries are short, we can safely make the assumption that a person will not be mentioned more than five times; thus, the model will look for features of the syntactic realizations from the first up to the fifth mention of an entity.

For each $m = 1, ..., 5$ we compute the probability

$$P(SF|m) = \frac{P(SF,m)}{P(m)} = \frac{\frac{cnt(SF,m)}{total}}{\frac{cnt(m)}{total}} = \frac{cnt(SF,m)}{cnt(m)},$$

where SF is the variable corresponding to the syntactic realization, m is the number of mention and $total$ is the number of syntactic forms counted for the entire corpus.

	first	second	third	foutrh	fifth
modified	0.76	0.48	0.52	0.54	0.51
non-modified	0.24	0.52	0.48	0.46	0.49
premodified	0.51	0.37	0.42	0.45	0.43
non-premodified	0.49	0.63	0.58	0.55	0.57
full name	0.97	0.13	0.12	0.10	0.10
last name	0.02	0.81	0.82	0.84	0.83
nickname	0.01	0.05	0.06	0.06	0.07

Figure 5.8: Probabilities of an NP being modified and non-modified at a particlular mention.

Model Comparison

The number of possible syntactic forms, which corresponds to the possible combination of features, is large, around 160. Because of this, it is not easy to interpret the results if they are taken in their full form. We now show information for one feature at a time so that the tendencies can become clearer.

Table 5.4 shows that a first mention is very likely to be modified in *some* way (probability of 0.76), but it is highly unlikely that it will be *both* premodified and postmodified (probability of 0.17).

The results for the presence or absence of modification of some kind are given in tables 5.4 and 5.8. The stratified model does not tell us anything about whether a subsequent mention should be modified; both cases are almost equally likely for mentions from the second up to the fifth. The Markov chain gives us more useful information, it predicts that at each next mention, modification can be either used or not, *but* once a non-modified form is chosen, the subsequent realizations will most likely not use modification any more.

The Markov chain that models the form of names also gives us more information than the stratified model. From the latter we can see that first name or nickname mentions are very unlikely. But the Markov chain also predicts that if such a reference is once chosen, it will most likely continue to be used as a form of reference. This is intuitively very appealing as it models cases where journalists call celebrities by their first name (e.g., "Britney" or

"Lady Diana" are commonly used while "Spears" or "Spencer" are not).

Also, the analysis of the data leads us to reject the hypothesis that there are some specific features of *each* number of mention. The distribution of features for mentions after the first are almost identical. This is one more reason to give preference to the Markov model over the stratified one, since the Markov model gives special importance to the first mention and treats all subsequent mentions equally.

Figure 5.6 shows the probabilities of transitions between the different kinds of postmodification. Prepositional, relative clause and "other" modifications appear with equal extremely low probability after any possible previous mention realization, even for initial reference. Thus the syntactic structure of the previous mention cannot be used as a predictor of the appearance of any of these kinds of modifications, so for the currently derived rules they will not be considered in any way but as "blockers" of further modification.

Figure 5.7 shows the probabilities for transitions between NPs with a different number of premodifiers. It can be seen that the mass above the diagonal is close to zero, which means that each subsequent mention has fewer premodifiers than the previous one. It is not surprising that a mention with one modifier is usually followed by a mention with one modifier (probability 0.5) since title modifiers such as "Mr." or "Mrs." are included in the counts for transitions. There are newspaper specific rules about the usage of these modifiers. The Wall Street Journal and the New York Times, for example, do use them as a rule, except for historical and criminal figures. Such editorial rules are interesting and can be useful for summarization, but the information that can be gathered from them is too subtle to encode computationally. As can be seen in the later section, these honorifics will be treated in rewrite as any other premodifier and will be dropped at subsequent mention.

We also looked separately at the cases when the first mention of a person was not the head of the noun phrase (e.g., "the Bush administration", "Mendeleev's periodic table"). Such name mentions obviously do not have postmodification of any kind and premodification cannot be reliably identified automatically since current parsers output flat structure for noun premodifiers. Words in the NP that precede the name can either modify the name or the head, so no conclusion can be drawn. The only relevant feature in this case was the name realization. The model built for those entities whose first mentions are in a non-head

NP differs quite a bit from the model for head NPs. This was the reason why in the rewrite rules that we developed and discuss below, the name is not changed in any way if its first mention is in a non-head position.

5.1.3 What was learned

The Markov chain model derived in the manner described above helps us understand what a typical text looks like. The Markov chain transitions give us defeasible preferences that are true for the average text. Human writers can be creative, so even statistically highly unlikely realizations can be used by a human writer. For example, even a first mention with a pronoun can be felicitous at times, as can be seen in figure 5.9. The fact that we were seeking preferences rather than rules allows us to take advantage of the sometimes inaccurate automatically derived corpus. There have inevitably been parser errors or mistakes in Nominator's output, but these can be ignored since, given the large amount of data, the general preferences in realization could be captured even from imperfect data.

He moved into the governor's mansion at 32, **heir to a tradition of progressive Southern governors** and ready to light up Arkansas. It was January 1979 and there was so much to do: Education needed to be overhauled, the business climate needed to be improved, the state needed to be dragged out of its slumberous, defeatist past. **Bill Clinton, the youngest governor in the nation since Harold Stassen,** had such big plans.

Figure 5.9: A first paragraph from our hand-annotated corpus. A first mention by pronoun is possible, but highly unlikely.

Since summaries are generated by a computer and not a human, deviation from the standard preference can very likely introduce a problem in the summary rather than make it more stylish. Thus the learned defeasible preferences help us decide when a reference in a summary needs to be rewritten and also it suggests the type of rewrite needed.

We developed a set of rewrite rules through manual analysis of the Markov chain model. This is a subset of the full power of the model, but it dramatically improves the quality of references. After we present the results for rule-based rewrite, we will also discuss how the limiting distribution of the Markov chain can be used to generate reference with more variability.

1. For first mentions
   (a) If the person's name is the head of the noun phrase,
      i. If a person is mentioned by last name, insert full name and the longest, in number of words, premodifier found in the input articles. The first mention from the article from which the summary sentence is drawn is preferred.
      ii. If no premodification is found in the input, check all first mentions in the input to see if any of them includes an apposition modifier. Take the longest such modifier and include it in the first mention NP.
   (b) The name is not modified at all if it is not the head of the noun phrase it appears in.
2. For all subsequent mentions use last name only, remove all premodifiers and delete all apposition modifiers.

The above straightforward rules lead to the following rewrite version of the summary in Figure 5.10.

Republican candidate Suzanne Haik Terrell had 56 percent of the white vote to 31 percent for **Democratic Sen. Mary Landrieu**, while **Landrieu** had 75 percent of the black vote to 10 percent for **Terrell**. A poll released this week shows the race between **Landrieu** and **her Republican challenger, Terrell,** to be dead even. Voters go to the polls Saturday. Emboldened by November election triumphs, President Bush urged Louisiana voters on Tuesday to pad the GOP Senate majority and defeat a Democratic incumbent who claims her own Bush-friendly voting record. With Louisiana's Senate run-off election just four days away, Bush led the GOP charge Tuesday for **Terrell** in what polls now suggest is a toss-up race against **Landrieu**.

Figure 5.10: Rewritten summary

5.1.4 Evaluation

The above three rules were used to rewrite 11 summaries chosen at random from the DUC 2001 and 2002 summaries that contained at least one reference to a person. Four human judges were then given the pairs of the original summary and its rewritten variant without being explicitly told which is which. They were asked to read the summaries and decide if they prefer one text over the other or if they are equal. They were also asked to give free-form comments on what they would change themsleves. The distribution of preferences is shown in figure 5.11.

In only one case a majority preference could not be reached, with two of the judges preferring the rewritten version and two, the original. This particular summary was controversial because it included non-name references to people, such as “the president” in the first coreference chain shown in figure 5.3 and not marked with a cross, indicating that it was not automatically recognized. This type of common noun coreference could not be identified in our automatic approach and thus, the fact of its occurrence was not taken into account during rewrite. This shows that work on person centered coreference can be very helpful for summarization as well.

There were two more cases where one judge showed preference for the original version. They both came with comments that the reason for the preference was that the original version exhibited more variation. Thus, it seems that the rule for strictly using last name at subsequent mentions is too rigid and most probably will need modification in cases where a person is mentioned more than three times.

rewrite version	**original version**	**none**
89%	9%	2%

Figure 5.11: Distribution of the 44 individual preferences for a rewritten or original summary.

5.1.5 Discussion and future work

We have shown how simple syntactic considerstions can improve a multi-document summary by making it more coherent. A Markov model for transitions between syntactic realizations was derived and used for composing initial rewrite rules. This approach to summarization, focusing on summary revision, has not been used in the area so far. Existing summarization approaches that do make any changes in the sentences from the original input have as a goal the reduction of information/number of words, while in our approach the coherence and readability of the summary are of primrary consideration.

As can be seen, a major improvement can be achieved even by using the proposed simple set of rewrite rules used for the evaluation. But they do not fully reflect all we learned from the data. These rules will be expanded with the rule for nickname usage discussed above. The rule for dropping premodification on subsequent mentions will also be refined so that it takes into account the gradual shrinking in the number of premodifiers. In order to do this we will need to build some kind of simple discourse model so that within it we can track which properties of an entity have already been realized and which can be realized in subsequent mentions.

One possible usage of the Markov model not discussed here is to use it to generate realizations "on demand" so that the highest probability path in the model can be realized in the summary. This means that referring expressions will be generated by recombining different pieces of the input rather then the currently used extraction of full NPs. For this task again a discourse model will be needed and the information in it will be used as a knowledge base for the generation process.

In order to use the Markov model directly for generation, we computed the limiting distribution of the chain. The limiting distribution gives the probability that the chain is in a given state at any time during transitions, regardless of the previous states it has been in. This is very convenient, since it does not involve conditional probabilities. The limiting distributions of form of the name, general modification, premodifiers and postmodifiers are given in the tables below.

Name form	Full name	Last name	First/Nickname
Probability	0.0838	0.8646	0.0517

Modification	Some modification	No modification
Probability	0.3043	0.6957

Number of premodifiers	0	1	2	3	4	5
Probability	0.7881	0.1562	0.0436	0.0100	0.0005	0.0016

Postmodifiers	Apposition	None	Prepositional	Relative clause	Other
Probability	0.0436	0.8862	0.0093	0.0321	0.0289

Now, the limiting distribution probabilities can be used to generate more variation for subsequent mentions (remember that in the rule-based approach subsequent mentions are always realized with the last name only). Let us take the name realization for example. At the point where subsequent mention needs to be generated, we produce a random number between 0 and 1. If the number is between 0 and 0.8646, the last name only is generated. If the number is between 0.8646 and 0.9484, the full name is generated (First name + Last name). And if the number is between 0.9484 and 1, just the first name is generated. Note that the limits for decision are taken from the limiting distribution of the name realization Markov chain. This generation process will lead to a more natural variation in subsequent mentions. Note that the above stationary distributions were dervived from a Markov chain trained on full documents rather than on summaries. As we discuss in the following section, some of the characteristics of references in summaries differ from references to the same entity in full news articles. If a large enough corpus of summaries is available, the model can be trained on summary data. The full evaluation of the procedure will be done in future work. We now turn to the discussion of cognitive status factors that can influence the form of first mentions in the summary.

5.2 Determining familiarity and importance of people in the news

To reiterate the discussion from the beginning of this chapter: extractive summaries contain phrases that the reader cannot understand out of context (Pai90) and irrelevant phrases that happen to occur in a relevant sentence (KM00; Bar03). Referring expressions in extractive summaries illustrate this problem, as sentences compiled from different documents might contain too little, too much or repeated information about the referent.

Whether a referring expression is appropriate depends on the location of the referent in the hearer's mental model of the discourse—the referent's *cognitive status* (GHZ93). If, for example, the referent is unknown to the reader at the point of mention in the discourse, the reference should include a description, while if the referent was known to the reader, no descriptive details are necessary.

Determining a referent's cognitive status, however, implies the need to model the intended audience of the summary. Can such a cognitive status model be inferred automatically for a general readership? We address this question by performing a study with human subjects to confirm that reasonable agreement on the distinctions can be achieved between different humans. We present an automatic approach for inferring what the typical reader is likely to know about people in the news. Our approach uses machine learning, exploiting features based on the form of references to people in the input news articles. Learning cognitive status of referents is necessary if we want to ultimately generate new, more appropriate references for news summaries and at the end of this chapter we demonstrate that the distinctions can indeed be used to reproduce human decisions on reference generation.

5.2.1 Hearer-Old vs Hearer-New

Hearer-new entities in a summary should be described in necessary detail, while hearer-old entities do not require an introductory description. This distinction can have a significant impact on overall length and intelligibility of the produced summaries. Usually, summaries are very short, 100 or 200 words, for input articles totaling 5,000 words or more. Several people might be involved in a story, which means that if all participants are fully described, little space will be devoted to actual news. In addition, introducing already familiar entities might distract the reader from the main story (Gri75). It is thus a good strategy to refer to an entity that can be assumed hearer-old by just a title + last name, e.g. *President Bush*, or by full name only, with no accompanying description, e.g. *Michael Jackson.*

5.2.2 Major vs Minor

Another distinction that human summarizers make is whether a character in a story is a major or a minor one and this distinction can be conveyed by using different forms of referring expressions. It is common to see in human summaries references such as *the dissident's father*. Usually, discourse-initial references solely by common noun, without the inclusion of the person's name, are employed when the person is not the main focus of a story (SMG88). By detecting the cognitive status of a character, we can decide whether to name the character in the summary. Furthermore, many summarization systems use the presence of named entities as a feature for computing the importance of a sentence or sentence fragment (SG04; GHW03; ZDL$^+$05). The ability to identify the major story characters and use only them for sentence weighting can benefit such systems since only 5% of all people mentioned in the input are also mentioned in the summaries.

5.2.3 Data preparation: the DUC corpus

The data we used to train classifiers for these two distinctions is the Document Understanding Conference collection (2001–2004) of 170 pairs of document input sets and the corresponding human-written multi-document summaries (2 or 4 per set). Our aim is to identify every person mentioned in the 10 news reports and the associated human summaries for each set, and assign labels for their cognitive status (hearer old/new and major/minor). To do this, we first preprocess the data and then perform the labeling and we discuss these steps in the next two sections.

Automatic preprocessing

All documents and summaries were tagged with BBN's IdentiFinder (BSW99) for named entities, and with a part-of-speech tagger and simplex noun-phrase chunker (GMMM00). In addition, for each named entity, relative clauses, appositional phrases and copula constructs, as well as pronominal co-reference were also automatically annotated (Sid03). We thus obtained coreference information (cf. Figure 5.12) for each person in each set, across documents and summaries.

Andrei Sakharov

Doc 1: [IR] laureate Andrei D. Sakharov [CO] Sakharov [CO] Sakharov [CO] Sakharov [CO] Sakharov [PR] his [CO] Sakharov [PR] his [CO] Sakharov [RC] who acted as an unofficial Kremlin envoy to the troubled Transcaucasian region last month [PR] he [PR] He [CO] Sakharov

Doc 2: [IR] Andrei Sakharov [AP] , 68 , a Nobel Peace Prize winner and a human rights activist , [CO] Sakharov [IS] a physicist [PR] his [CO] Sakharov

Figure 5.12: Example information collected for *Andrei Sakharov* from two news report. 'IR' stands for 'initial reference', 'CO' for noun co-reference, 'PR' for pronoun reference, 'AP' for apposition, 'RC' for relative clause and 'IS' for copula constructs.

The tools that we used were originally developed for processing single documents and we had to adapt them for use in a multi-document setting. The goal was to find, for each person mentioned in an input set, the list of all references to the person in both input documents and human summaries. For this purpose, all input documents were concatenated and processed with IdentiFinder. This was then automatically post-processed to markup coreferring names and to assign a unique canonical name (unique id) for each name coreference chain. For the coreference, a simple rule of matching the last name was used, and the canonical name was the "FirstName LastName" string where the two parts of the name could be identified [3]. Concatenating all documents assures that the same canonical name will be assigned to all named references to the same person.

The tools for pronoun coreference and clause and apposition identification and attachment were run separately on each document. Then the last name of each of the canonical names derived from the IdentiFinder output was matched with the initial reference in the generic coreference list for the document with the last name. The tools that we used have been evaluated separately when used in normal single document setting. In our cross-

[3] Occasionally, two or more different people with the same last name are discussed in the same set and this algorithm would lead to errors in such cases. We did keep a list of first names associated with the entity, so a more refined matching model could be developed, but this was not the focus of this work.

document matching processes, we could incur more errors, for example when the general coreference chain is not accurate. On average, out of 27 unique people per cluster identified by IdentiFinder, 4 people and the information about them are lost in the matching step for a variety of reasons such as errors in the clause identifier, or the coreference.

Data labeling

Entities were automatically labeled as hearer-old or new by analyzing the syntactic form that human summarizers used for initial references to them. The labeling rests on the assumption that the people who produced the summaries used their own model of the reader when choosing appropriate references for the summary. The following instructions had been given to the human summarizers, who were not professional journalists: "To write this summary, assume you have been given a set of stories on a news topic and that your job is to summarize them for the general news sections of the Washington Post. Your audience is the educated adult American reader with varied interests and background in current and recent events." Thus, the human summarizers were given the freedom to use their assumptions about what entities would be generally hearer-old and they could refer to these entities using short forms such as (1) title or role+ last name or (2) full name only with no pre- or post-modification. Entities that the majority of human summarizers for the set referred to using form (1) or (2) were labeled as hearer-old. From the people mentioned in human summaries, we obtained 118 examples of hearer-old and 140 examples of hearer-new persons, 258 examples in total, for supervised machine learning.

In order to label an entity as major or minor, we again used the human summaries—entities that were mentioned *by name* in at least one summary were labeled *major*, while those not mentioned by name in any summary were labeled *minor*. The underlying assumption is that people who are not mentioned in any human summary, or are mentioned without being named, are not important. There were 258 major characters who made it to a human summary and 3926 minor ones that only appeared in the news reports. Such distribution between the two classes is intuitively plausible, since many people in news articles express opinions, make statements or are in some other way indirectly related to the story, while there are only a few main characters.

0,1:	Number of references to the person, including pronouns (total and normalized by feature 16)	2,3:	Number of times apposition was used to describe the person(total and normalized by feature 16)
4,5:	Number of times a relative clause was used to describe the person (total and normalized by 16)	6:	Number of times the entity was referred to by name after the first reference
7,8:	Number of copula constructions involving the person (total and normalized by feature 16)	9,10:	Number of apposition, relative clause or copula descriptions (total and normalized by feature 16)
11,12,13:	Probability of an initial reference according to the bigram model (av.,max and min of all initial references)	14:	Number of top 20 high frequency description words (from references to people in large news corpus) present in initial references
15:	Proportion of first references containing full name	16:	Total number of documents containing the person
17,18:	Number of appositives or relative clause attaching to initial references (total and normalized by feature 16)		

Table 5.1: List of Features provided to WEKA.

5.2.4 Machine learning experiments

For our experiments, we used the Weka (WF05) machine learning toolkit and obtained the best results for hearer-old/new using a support vector machine (SMO algorithm) and for major/minor, a tree-based classifier (J48). We used Weka's default settings for both algorithms.

We now discuss what features we used for our two classification tasks (cf. list of features in table 5.1). Our hypothesis is that features capturing the frequency and syntactic and lexical forms of references are sufficient to infer the desired cognitive model.

Intuitively, pronominalization indicates that an entity was particularly salient at a specific point of the discourse, as has been widely discussed in attentional status and centering literature (GS86; GGG93). Modified noun phrases (with apposition, relative clauses or premodification) can also signal different status.

In addition to the syntactic form features, we used two months worth of news articles collected over the web (and independent of the DUC collection we use in our experiments here) to collect unigram and bigram lexical models of first mentions of people. The names themselves were removed from the first mention noun phrase and the counts were collected over the premodifiers only. One of the lexical features we used is whether a person's description contains any of the 20 most frequent description words from our web corpus. We reasoned that these frequent descriptors may signal importance; the full list is:

> *president, former, spokesman, sen, dr, chief, coach, attorney, minister, director, gov, rep, leader, secretary, rev, judge, US, general, manager, chairman.*

Another lexical feature was the overall likelihood of a person's description using the bigram model from our web corpus. This indicates whether a person has a role or affiliation that is frequently mentioned. We performed 20-fold cross validation for both classification tasks. The results are shown in Table 5.2 (accuracy) and Table 5.3 (precision/recall).

5.2.5 Major vs. Minor results

For major/minor classification, the majority class prediction has 94% accuracy, but is not a useful baseline as it predicts that *no* person should be mentioned by name and all are minor characters. J48 correctly predicts 114 major characters out of 258 in the 170 document sets. As recall appeared low, we further analyzed the 148 persons from DUC'03 and DUC'04 sets, for which DUC provides four human summaries. Table 5.4 presents the distribution of recall taking into account *how many* humans mentioned the person by name in their summary (originally, entities were labeled as main if *any* summary had a reference to them, cf. §16). It can be seen that recall is high (0.84) when all four humans consider a character to be major, and falls to 0.2 when only one out of four humans does. These observations reflect the well-known fact that humans differ in their choices for content selection, and indicate that in the automatic learning is more successful when there is more human agreement.

In our data there were 258 people mentioned by name in at least one human summary. In addition, there were 103 people who were mentioned in at least one human summary using only a common noun reference (these were identified by hand, as common noun coreference cannot be performed reliably enough by automatic means), indicating that 29% of people mentioned in human summaries are not actually named. Examples of such references include *an off duty black policeman*, *a Nigerian born Roman catholic priest*, *Kuwait's US ambassador*. For the purpose of generating references in a summary, it is important to evaluate how many of these people are correctly classified as minor characters. We removed these people from the training data and kept them as a test set. WEKA achieved a testing accuracy of 74% on these 103 test examples. But as discussed before, different human summarizers sometimes made different decisions on the form of reference to use. Out of the 103 referents for which a non-named reference was used by a summarizer, there were 40 where other summarizers used named reference. Only 22 of these 40 were labeled as minor characters in our automatic procedure. Out of the 63 people who were not named in *any* summary, but mentioned in at least one by common noun reference, WEKA correctly predicted 58 (92%) as minor characters. As before, we observe that when human summarizers generate references of the same form (reflecting consensus on conveying the perceived importance of the character), the machine predictions are accurate.

We performed feature selection to identify the most important features for the classification task. For the major/minor classification, the important features used by the classifier were the number of documents the person was mentioned in (feature 16), number of mentions within the document set (features 1,6), number of relative clauses (feature 4,5) and copula (feature 8) constructs, total number of apposition, relative clauses and copula (feature 9), number of high frequency premodifiers (feature 14) and the maximum bigram probability (feature 12). It was interesting that presence of apposition did not select for either major or minor class. It is not surprising that the frequency of mention within and across documents were significant features—a frequently mentioned entity will naturally be considered important for the news report. Interestingly, the syntactic form of the references was also a significant indicator, suggesting that the centrality of the character was signaled by the journalists by using specific syntactic constructs in the references.

	Major/Minor	Hearer New/Old
Weka	0.96 (J48)	0.76 (SMO)
Majority class prediction	0.94	0.54

Table 5.2: Cross validation *testing* accuracy results.

	Class	Precision	Recall	F-measure
SMO	hearer-new	0.84	0.68	0.75
	hearer-old	0.69	0.85	0.76
J48	major-character	0.85	0.44	0.58
	minor-character	0.96	0.99	0.98

Table 5.3: Cross validation *testing* P/R/F results.

Hearer Old vs New Results

The majority class prediction for the hearer-old/new classification task is that no one is known to the reader and it leads to overall classification accuracy of 54%. Using this prediction in a summarizer would result in excessive detail in referring expressions and a consequent reduction in space available to summarize the news events. The SMO prediction outperformed the baseline accuracy by 22% and is more meaningful for real tasks.

For the hearer-old/new classification, the feature selection step chose the following features: the number of appositions (features 2,3) and relative clauses (feature 5), number of mentions within the document set (features 0,1), total number of apposition, relative clauses and copula (feature 10), number of high frequency premodifiers (feature 14) and the minimum bigram probability (feature 13). As in the minor-major classification, the syntactic choices for reference realization were useful features.

We conducted an additional experiment to see how the hearer old/new status impacts the use of apposition or relative clauses for elaboration in references produced in human summaries. It has been observed (SNM04) that on average these constructs occur 2.3 times *less* frequently in human summaries than in machine summaries. As we show, the use of postmodification to elaborate relates to the hearer-old/new distinction.

To determine when an appositive or relative clause can be used to modify a reference, we considered the 151 examples out of 258 where there was at least one relative clause

Number of summaries containing the person	Number of examples	Number and % recalled by J48
1 out of 4	59	15 (20%)
2 out of 4	35	20 (57%)
3 out of 4	29	23 (79%)
4 out of 4	25	21 (84%)

Table 5.4: J48 Recall results and human agreement.

or apposition describing the person in the input. We labeled an example as positive if *at least* one human summary contained an apposition or relative clause for that person and negative otherwise. There were 66 positive and 85 negative examples. This data was interesting because while for the majority of examples (56%) all the human summarizers agreed not to use postmodification, there were very few examples (under 5%) where all the humans agreed to postmodify. Thus it appears that for around half the cases, it should be obvious that no postmodification is required, but for the other half, human decisions go either way.

Notably, none of the hearer-old persons (using test predictions of SMO) were postmodified. Our cognitive status predictions cleanly partition the examples into those where postmodification is not required, and those where it might be. Since no intuitive rule handled the remaining examples, we added the testing predictions of hearer-old/new and major/minor as features to the list in Table 5.1, and tried to learn this task using the tree-based learner J48. We report a testing accuracy of 71.5% (majority class baseline is 56%). There were only three useful features—the predicted hearer-new/old status, the number of high frequency premodifiers for that person in the input (feature 14 in table 5.1) and the average number of postmodified initial references in the input documents (feature 17).

5.2.6 Validating the results on current news

We tested the classifiers on data different from that provided by DUC, and also tested human consensus on the hearer-new/old distinction. For these purposes, we downloaded 45 clusters from one day's output from Newsblaster[4]. We then automatically compiled the list

[4] `http://newsblaster.cs.columbia.edu`

of people mentioned in the machine summaries for these clusters. There were 107 unique people that appeared in the machine summaries, out of 1075 people in the input clusters.

Human agreement on hearer-old/new

The distinction between hearer-old and hearer-new entities depends on the readers. In other words, we are attempting to automatically infer which characters would be hearer-old *for the intended readership of the original reports*, which is also expected to be the intended readership of the summaries. A question arises when attempting to infer hearer-new/old status: Is it meaningful to generalize this across readers, seeing how dependent it is on the world knowledge of individual readers?

To address this question, we gave four American graduate students (all male, born in the US and native speakers of English) a list of the names of people in the DUC human summaries (cf. §5.2.3), and asked them to write down for each person, their country/state/organization affiliation and their role (writer/president/attorney-general etc.). We considered a person hearer-old to a subject if they correctly identified both role and affiliation for that person. For the 258 people in the DUC summaries, the four subjects demonstrated 87% agreement ($\kappa = 0.74$)[5]. As reported in the section of machine learning results, 54% of the entities were hearer-new.

Similarly, they were asked to perform the same task for the Newsblaster data, which dealt with contemporary news[6], in contrast with the DUC data that contained news from the late 80s and early 90s. On this data, 80% of the entities were hearer-new and the human agreement was 91% ($\kappa = 0.78$). This is a high enough agreement to suggest that the classification of national and international figures as hearer old/new across *the educated adult American reader with varied interests and background in current and recent events* is a well defined task. This is not necessarily true for the full range of cognitive status distinctions; for example Poesio and Viera (PV98) report lower human agreement on more

[5] κ (kappa) is a measure of inter-annotator agreement over and above what might be expected by pure chance (See Carletta (Car96) for discussion of its use in NLP). $\kappa = 1$ if there is perfect agreement between annotators and $\kappa = 0$ if the annotators agree only as much as you would expect by chance.

[6] The human judgments were made within a week of the news stories appearing.

fine-grained classifications of definite descriptions.

Results on the Newsblaster data

We measured how well the models trained on DUC data perform with current news labeled using human judgments. For each person who was mentioned in the automatic summaries for the Newsblaster data, we compiled one judgment from the four human subjects: an example was labeled as hearer-new if two or more out of the four subjects had marked it as hearer new. Then we used this data as *test data*, to test the model trained solely on the DUC data. The classifier for hearer-old/hearer-new distinction achieved 75% accuracy on Newsblaster data labeled by humans, while the cross-validation accuracy on the automatically labeled DUC data was 76%. These numbers are very encouraging, since they indicate that the performance of the classifier is stable and does not vary between the DUC and Newsblaster data. The precision and recall for the Newsblaster data are also very similar for those obtained from cross-validation on the DUC data:

Class	Precision	Recall	F-Measure
Hearer-old	0.88	0.73	0.80
Hearer-new	0.57	0.79	0.66

Major/Minor results on Newsblaster data

For the Newsblaster data, no human summaries were available, so no direct indication on whether a human summarizer will mention a person in a summary was available. In order to evaluate the performance of the classifier, we gave to another graduate student the list of people's names appearing in the machine summaries, together with the input cluster and the machine summary, and asked which of the names on the list would be a suitable keyword for the set (keyword lists are a form of a very short summary). Out of the 107 names on the list, the annotator chose 42 as suitable for descriptive keyword for the set.

The major/minor classifier was run on the 107 examples; only 40 were predicted to be major characters. Of the 67 test cases that were predicted by the classifier to be minor characters, 12 (18%) were marked by the annotator as acceptable keywords. In comparison, of the 40 characters that were predicted to be major characters by the classifier, 30

(75%) were marked as possible keywords. If the keyword selections of the annotator are taken as ground truth, the automatic predictions have precision and recall of 0.75 and 0.71 respectively for the *major* class.

5.2.7 Selecting appropriate first mention

In this section, present experiments that demonstrate how the hearer new/old classification we have derived can be used in conjunction with a discourse model to address the issue of what pre- and post-modification to generate for initial references to people in summaries.

An analysis of premodification in initial references to people in DUC human summaries showed that 71% of premodifying words were either title or role words (eg. *Prime Minister, Physicist* or *Dr.*) or reference modifying adjectives such as *former* that have to be included with the role. Due to journalistic conventions (in the context of DUC, the human summarizer tended to follow the conventions in the source news reports), the initial references to almost all persons included such a role or title premodification in human summaries. Indeed a simple rule – to include the role, where available, in initial references – reproduced the choices made by the human summarizers with only a handful of exceptions. We test the validity of our selection rules by comparing their predictions to the decisions made by DUC summarizers, we need to include role. We can nonetheless postulate that for greater compression, the role or title can be omitted for hearer-old persons; for example generating *Margaret Thatcher* instead of *Former Prime Minister Margaret Thatcher.*

Thus the most important generation decision left to be made regarding premodification is when to include affiliation — country, state or organization names constituted 22% of premodifying words. All other kinds of premodifying words, such as *moderate, celebrity* and *loyal* constitute only 7%.

We now describe a procedure that uses hearer and discourse information to decide when to provide an affiliation in the initial reference to a person. This issue is ubiquitous in summarizing news; for example, the reference generator might need to decide between *White House Press Secretary James Brady* and *Press Secretary James Brady*, between *Soviet President Gorbachev* and *President Gorbachev* or between *Indiana Senator Dan Quayle* and *Senator Dan Quayle.*

The Decision Procedure

Based on our intuitions about discourse salience and information status, we initially postulated the following decision procedure:

Rule 1 The affiliation of a person can be omitted in the first-mention reference if:

(a) The person is classified as hearer-old.

(b) Or, if the person's organization (country/ state/ affiiation) has been already mentioned and is the most salient organization in the discourse at the point where the reference needs to be generated.

We described how we make the hearer new/old judgment in §5.2.4. We used a salience-list (S-List) (Str98) to determine the salience of organizations. This is a shallow attentional-state model and works as follows:

1. Within a sentence, entities are added to the salience-list from left to right.
2. Within the discourse, sentences are considered from right to left.

In other words, entities in more recent sentences are more salient than those in previous ones and within a sentence, earlier references are more salient than later ones.

Results

To make the evaluation non-trivial, we only considered examples where there was an affiliation mentioned for the person in the input documents, ruling out the trivial cases where there was no choice to be made. There were 272 initial references to 182 persons in the human summaries that met this criterion (note that there were multiple human summaries for each document set).

We used 139 of these 272 examples (from DUC'01, '02 and '03) as training data to check and possibly refine our rule. For each of these 139 initial references to people, we:

1. Obtained from the source news reports the test-set prediction from WEKA on whether that person was hearer-new or hearer-old.
2. Formed the S-List for affiliations in that human summary at the point of reference[7].

[7] *http://www.cia.gov/cia/publications/factbook* provides a list of countries and states, abbreviations and adjectival forms, while the named entity recognition tool IDENTIFINDER marks up organizations. The output was manually cleaned to remove errors in Named Entity detection.

3. Used the decision procedure described above to decide whether or not to include the affiliation in the reference.

The evaluation consisted of matching our predictions with the observed references in the human summaries. Our decision procedure made the correct decision in 71% of the instances and successfully modelled variations in the initial references used by different human summarizers for the same document set:

1. **Brazilian President Fernando Henrique Cardoso** was re-elected in the...
   [*hearer new* and Brazil not in context]
2. Brazil's economic woes dominated the political scene as **President Cardoso**...
   [*hearer new* and Brazil most salient country in context]

and in the initial reference to the same person across summaries of different document sets:

1. It appeared that **Iraq's President Saddam Hussein** was determined to solve his countries financial problems and territorial ambitions...
   [*hearer new* for this document set and Iraq not in context]
2. ...A United States aircraft battle group moved into the Arabian Sea. **Saddam Hussein** warned the Iraqi populace that United States might attack...
   [*hearer old* for this document set]

An error analysis showed that in most of these instances the rule predicted no affiliation in instances where the human summarizer had included it. In many cases, the person was first mentioned in a context where a different organization/state or country was more salient than their own. When we modified condition (1) of our decision rule 1 to obtain:

Rule 2 The affiliation of a person can be omitted in the first-mention reference if:

(a) The person is hearer-old, *and no country/state/org is more salient than their own.*

(b) Or, if the person's affiliation has been already mentioned and is the most salient of all the organizations in the discourse at the point where the reference needs to be generated.

Rule 2 increased the accuracy to 78%. The improved performance of our second decision procedure suggests that affiliation is sometimes included in references to even hearer-old persons in order to aid the hearer in immediately recollecting the referent. However both algorithms make errors on such cases, and there appears to be some variability in how human summarizers make their decisions in these contexts.

Having convinced ourselves about the validity of these rules, we applied them to the 133 examples in the unseen test data. The results are shown in table 5.5 below. 85% of the observed human references were modelled correctly by either Rule 1 or Rule 2. The remaining errors were largely due to misclassifications of people as hearer new by SMO, thus leading our rule to include affiliation when not required. We compared the rules' prediction accuracy to that of three baselines (see table 5.5):

1. **Never-Include**: This is the majority class baseline which says that affiliation is always omitted.
2. **Information-Status**: Always include if hearer-new, never include if hearer old (using testing predictions from automatic classification of information status).
3. **Salience**: Include affiliation unless that affiliation is already most salient at the point of reference.

Algorithm	Accuracy
Never-Include Baseline	0.56
Information-Status Baseline	0.58
Salience Baseline	0.65
Salience+Information Status (Rule 1)	0.79
Salience+Information Status (Rule 2)	0.75

Table 5.5: Test set results for decision procedure to include affiliation in initial references.

The two rules that we introduced for generation of first mention outperform all three baselines, by as much as 14% increase in accuracy.

5.3 Conclusions

Cognitive status distinctions are important when generating summaries, as they help determine both what to say and how to say it. However, to date, no one has attempted the task of inferring cognitive status from unrestricted news.

We have shown that the hearer-old/new and major/minor distinctions can be inferred using features derived from the lexical and syntactic forms and frequencies of references in the news reports. We have presented results that show agreement on the *familiarity* distinction between educated adult American readers with an interest in current affairs, and that the learned classifier accurately predicts this distinction. We have demonstrated that the acquired cognitive status is useful for determining which characters to name in summaries, and which named characters to describe or elaborate. In addition, we evaluated a rule-based approach to rewrite of references to people and of generation of first time reference that led to results much better than the baseline.

This provides the foundation for a principled framework in which to address the question of how much references can be shortened without compromising readability.

Chapter 6

Final discussion: contributions and limitations

This thesis is concerned with the investigation and understanding of the process of generic multi-document summarization, including the study of the level of human agreement on content selection, system evaluation, and aspects of system development.

The main contributions of the thesis include:

Context sensitive frequency-based summarizer We built a family of frequency-based summarizers. Three main steps of the summarizer were proposed: *1)* Assign weights to content words. *2)* Choose a composition function to combine the weights of words into weights of sentences. *3)* After the choice of each sentence, make explicit context sensitive adjustment of the weights of words. We demonstrated that each of these steps can have a significant impact on the performance of the summarizer. By avoiding formal and theoretical complications we have shown *which are the steps* in a summarizer that matter for good performance. We showed that the context sensitive summarizer with *Average* as a combination function achieves performance as good as the state-of-the-art summarizer. This is a solid achievement, since our summarizer, unlike the state-of-the-art, does not need training data, background corpora, and does not rely on ad-hoc sets weights or parameters. The insights gained from our experiments now allow us to look for better formal models to capture the intuitions about frequency and context.

Summary rewrite In this thesis, we introduced the concept of summary rewrite of maximum noun phrases and of references to people.

- Our approach to noun phrase rewrite demonstrated how the combination of importance and context considerations can be used on subsentential units. For the noun phrase rewrite, we use the redundancy in the input to find the noun phrases that are most suitable for *the current context* of the summary. The method led to summaries that were 50% different from the original extractive summaries. The rewritten summaries had better scores for content selection in half of test cases, and led to marginal overall improvement of content selection. The noun phrase rewrite approach produced summaries that were significantly better in grammaticality, clarity of reference and coherence than an event-based generative summarizer.
- For the rewrite of references to people, we explored two distinctions. For first mention vs. subsequent mention, we used a Markov chain model to capture the appropriate syntactic form of reference. We showed that humans prefer summaries in which the first mention to a person includes a description and subsequent mentions are short. For first mentions, we further explored possible difference of the reference depending on its assumed familiarity to the reader. Our rule-based approach achieved 80% accuracy in reproducing human decision for first reference, outperforming salience and majority baselines.

Automatic learning of cognitive status distinctions We developed classifiers based on shallow lexical and syntactic features, that decide if an entity mention in the input to the summarizer is *major* or *minor*, that is if it is globally salient and should be included in a summary or not. Another classifier automatically decides if entities from the input are *hearer-old* or *hearer-new*, that is, if the intended readers of the input are likely to be already familiar with the entity or not. The acquisition of such information about entities have not been explored in the past. We demonstrated that these distinctions can be used to reproduce human choices for generation of first mentions to people in our experiments with rewrite of references to people.

Pyramid evaluation method We developed an annotation scheme to highlight similarities and differences between several texts on the subsentential level. The method not only allowed for a better study of human agreement in content selection, but also served as a foundation of an empirically sound evaluation model. The method provides greater reward for content that several humans would agree to include in their summaries. The method has satisfied the need in the summarization community for an evaluation approach that uses multiple models, avoiding the bias a single model can inpose on the results.

Study of frequency in the input as an indicator of importance Generic multi-document summarization can enhance news-browsing sites, and much progress has been made in the summarization field, and yet, little was known about what makes automatic summarizers good prior to our study. Our empirical study focused on one feature used as indicator if importance: frequency in the input. We observed that for both content words and content units, higher frequency in the input is predictive of the word or content unit appearing in a human summary, and that human summarizers tend to agree on the inclusion of content that is frequently repeated in the input. Human summaries also have higher likelihood under a multinomial model estimated from the input. We have thus empirically shown that frequency is a good feature for single document summarization. We hope that our approach will be used in the future to validate the use of other features for summarization.

6.1 Limitations and future work

In this thesis we demonstrated that most of the successful automatic summarizers for generic multi-document summarization of news need a a means to account for context and a good composition function to assign weights to larger text chunks such as sentences. The natural question that arises is whether the same features and techniques will be useful for other types of summarization, for example for the creation of update summaries, or query-focused summaries. In the thesis, we explored frequency in the input as a feature to assign importance to basic content units: content words. Such a feature is unlikely to be helpful in these different, more difficult, types of summarization for news. Nonetheless, even for these new tasks, we expect context and the choice of composition function to be very important. We outlined a principled approach to the study of the problem of identifying useful features through empirical comparison between characteristics of the input and the observed human agreement in content selection: in the future we could study how predictive suggested features are, based on their manifestation in human summaries.

We would also like to explore more formal models that can capture the intuitive steps we included in the frequency-based summarizer. We used a standard maximum likelihood estimator for the probability of words in the input. But there are other models, developed to better capture the Zipfian distribution of words and content units. Several of these are discussed for example in (Baa01). For example, a lognormal model has been applied to word frequency distributions by viewing the use of a word as resulting from a selection process through a binary branching tree. Decision probabilities are assigned to each branching point in the decision tree. The individual words of the vocabulary appear at the leaf nodes. Each word is associated with a unique path in the tree. The probability of selecting a word is the product of the decision probabilities of its path. Such more sophisticated models could lead to better results, capturing better human performance. The exploration of such models is part of our future work.

The approach to summary rewrite also merits future research. In the experiments presented here, generic noun phrase rewrite lead to deterioration of the linguistics aspects of the summaries compared to extractive summarization, notably the grammaticality and the referential clarity of references suffered. We would like to further explore abstractive strategies that combine grammaticality concerns with constituent importance. We hope that stand-alone generation components such as those developed in NLPWin can be used to achieve better results for summary rewrite.

We would also like to integrate our cognitive status classification with online summarization. For example, the classifiers can be used to identify people mentioned in the automatic summaries that are major characters and unknown to the readers. Additional “Who is X” summaries can be automatically generated by a question answering system for such people, providing a link to the full person description.

Finally, throughout the thesis we discussed the readability problems typical for current summarization systems. More than half of the automatic summaries were judged as barely acceptable to poor on qualities such as clarity or reference, focus and coherence. An issue that was not discussed in the thesis and needs to receive further attention is that of developing models of focus and coherence. For example, the context sensitivity in the frequency-based summarizer indicates which entities or events *do not need to be mentioned* any more in the summary. A good model would be able to predict *what content is suitable to follow next.* We plan to address such problems in future research. The pyramid annotation of human summaries can help develop such models. In human summaries, we see what content has been chosen, and how it has been organized in the summaries. The availability of several human summaries allows to investigate the strength of the information grouping and ordering constraints.

6.2 Final discussion

We will now reiterate several points made in thesis.

Need for complete evaluation Traditionally, content selection and linguistic quality have been evaluated separately in summarization. In fact, linguistic quality has often been overlooked, with researchers focusing on publishing results only on content selection. Our experiments with NP-rewrite show the danger of such an approach. In terms of content selection, generic NP-rewrite led to modest improvements. But it also led to major deterioration of grammaticality, clarity of reference and coherence. If the results in content selection were even a bit better, one could have claimed improvements. But in order to assess the full impact of a technique, one needs to evaluate *both* aspects of the produced summaries, as we did in this thesis, to be able to draw conclusions that reflect the performance of the summarizer as a whole.

Manual and automatic evaluation The use of automatic evaluation metrics is rather tempting, given their low cost and high speed. Our experiments with the frequency-based summarizer family showed that automatic metrics can be used to suggest further steps, and to narrow down the choices for manual evaluation, but overall the manual evaluation remains the only trustworthy method for making final decisions on system performance. Specifically, we observed that the unigram metric, ROUGE-1, with stopwords removed, led to conclusions that were closest to those we drew based on manual evaluation. While the unigram match metric might seem too simplistic and undesirable, the other two popular metrics (bigram and skip bigram) would have led us to make incorrect conclusions, claiming that the frequency-based summarizer significantly outperforms the state-of-the-art system. We did see that the empirical distribution of content words and content units in a pool of human summaries for the same input are very similar, indicating that the overall good correlation between the unigram overlap method and the manual methods is not due to chance.

In the thesis we showed that frequency in the input is highly predictive. This fact leads to the question: can frequency in the input be directly used to assign weights to content units, making the creation of human summaries unnecessary for evaluation? In fact in previous research in indicative generic single document summarization, Donaway *et al.* have suggested that the creation of ground truth summaries is unnecessary and a summary can be evaluated through comparison with the input (DDM00). To address the plausibility of this suggestion in the context of multi-document summarization, we evaluated an automatic system using the pyramid method on the 11 sets, using both a pyramid directly derived from the input documents and a pyramid built from human summaries as the original pyramid method prescribes. We calculated the correlation between the scores assigned by the two methods for the 11 summaries. Pearson's correlation coefficient was 0.83 (p-value= 0.0103) and Spearman's correlation was 0.81 (p-value= 0.0348). Both correlations are significant at the 95% level of significance, but not high enough to be considered mutually substitutable. This means that direct content unit annotation of the input for a summarizer cannot be used directly for evaluation.

Abstraction techniques In parts of this thesis we explored generation approaches for summarization. We believe such techniques are necessary in order to achieve close to human performance. For example, in the beginning of the thesis, we mentioned that in single-document summarization, systems do not outperform the first paragraph baseline, while human summarizers significantly outperform the baseline. In multi-document summarization, we saw that a single feature, frequency, suffices to reach above the baseline performance. It is possible that in single document summarization extractive systems do not do so well since there is not enough redundancy (repetition or frequency) to estimate importance. We conjecture that in single document summarization, information reduction and information combination, non-extractive techniques, are the only way to achieve better performance.

But especially in the beginning stages of development of abstractive techniques, it is likely they will introduce grammatical errors and will perform poorly when compared with extractive methods. For this purpose, it seems that the best solution is to create a special generation task for summarization, in which systems need to demonstrate that they change a certain percentage of an extractive summary in order to be entered in the track. Such a set-up will allow more freedom for exploration of generation techniques and can help bring better results in the future and it will not stifle inovative research under the unfair evaluation pressure.

References

[ABN00] R. Ando, B. Boguraev, and M. Neff. Multi-document summarization by visualizing topical content. In *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization*, pages 79–88, 2000.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[Baa01] R. Harald Baayen. *Word Frequency Distributions.* Kluwer Academic Publishers, 2001.

[Bar03] Regina Barzilay. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation.* PhD thesis, Columbia University, New York, 2003.

[BE03] Regina Barzilay and Noemie Elhadad. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, 2003.

[BEM02] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring strategies for sentence ordering in multi-document summar ization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.

[BKN01] Endre Boros, Paul Kantor, and David Neu. A clustering based approach to creating multi-document summaries. In *Proceedings of the 1st Document Understanding Conference (DUC'01)*, 2001.

[BM00] Jill Burstein and Daniel Marcu. Toward using text summarization for essay-based feedback. In *Proceedings of TALN 2000 Conference*, 2000.

[BM05] Regina Barzilay and Kathleen McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3), 2005.

[BME99] Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[BMR95] Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[BN00] B. Boguraev and M. Neff. Lexical cohesion, discourse segmentation and document summarization. In *RIAO-2000, Content-Based Multimedia Information Access*, 2000.

[BSW99] D. Bikel, R. Schwartz, and R. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231, 1999.

[BV04] Michele Banko and Lucy Vanderwende. Using n-grams to understand the nature of summaries. In *Proceedings of HLT/NAACL'04*, 2004.

[Car96] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[CDS96] Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. Motivation and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 1041–1044, 1996.

[CG98] Jaime Carbonell and Jade Goldstein. 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 (SIGIR'98)*, pages 335–336, 1998.

[Cha00] Eugene Charniak. A maximum-entropy-inspired parser. In *NAACL-2000*, 2000.

[Cha01] Eugene Charniak. Unsupervised learning of name structure from coreference data. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 48–54, 2001.

[CMP$^+$99] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devin, and John Tait. Simplifying english text for language impaired readers. In *Proceedings of the 9th Conference of the Eurpoean Chapter of the Association for Computational Linguistics (EACL'99)*, pages 269–270, 1999.

[CS04] Terry Copeck and Stan Szpakowicz. Vocabulary agreement among model summaries and source documents. In *Proceedings of the Document Understanding Conference DUC'04*, 2004.

[CSGO04] John Conroy, Judith Schlesinger, Jade Goldstein, and Dianne O'Leary. Left-brain/right-brain multi-document summarization. In *Proceedings of the 4th Document Undersatnding Conference (DUC'04)*, 2004.

[CSOO01] John Conroy, Judith Schlesinger, Dianne O'Leary, and Mary Ellen Okurowski. Using hmm and logistic regression to generate extract summaries for duc. In *Proccedings of the 1st Document Understanding Conference (DUC'01*, 2001.

[CW05] Richard Craggs and Mary McGee Wood. Evaluating discourse and dialog coding schemes. *Computational Linguistics*, 31(3):289–295, 2005.

[DCS+03] Daniel Dunlavy, John Conroy, Judith Schlesinger, Sarah Goodman, Mary Ellen Okurowski, and Dianne O'Leary. Performance of a three-stage system for multi-document summarization. In *Proceedings of the 3rd Document Understanding Conference (DUC'03)*, 2003.

[DDM00] Robert Donaway, Kevin Drummey, and Laura Mather. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of NAACL-ANLP Workshop on Automatic Summarization*, 2000.

[DG04] Barbara DiEugenio and Michael Glass. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 2004.

[DM69] Wilfrid Dixon and Frank Massey. *Introduction to statistical analysis.* McGraw-Hill Book Company, 1969.

[DM04] Hal Daumé III and Daniel Marcu. A phrase-based HMM approach to document/abstract alignment. In *Proceedings of EMNLP*, 2004.

[Dun94] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1994.

[EKKM05] Noemie Elhadad, Min-Yen Kan, Judith Klavans, and Kathleen McKeown. Customization in a unified framework for summarizing medical literature. *Journal of Artificial Intelligence in Medicine*, 33:179–198, 2005.

[EM05] David Kirk Evans and Kathleen McKeown. Identifying similarities and differences across english and arabic news. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[ER04a] Güneş Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.

[ER04b] Gunes Erkan and Dragomir Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.

[Fox98] B. Fox. *Discourse structure and anaphora.* Cambridge University Press, 1998.

[Fra90] K. Fraurud. Definiteness and the processing of nps in discourse. *Journal of Semantics*, 7:395–433, 1990.

[GA04] Chung Heong Gooi and James Allan. Cross-document coreference on a large scale corpus. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[Gar82] Ruth Garner. Efficient text summarization. *Journal of Educational Research*, 75:275–279, 1982.

[GG04] R. V. Guha and A. Grag. Disambiguating people in search. In *Proceedings of the 13th World Wide Web conference (WWW 2004)*, 2004.

[GGG93] P. Gordon, B. Grosz, and L. Gilliom. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347, 1993.

[GHW03] Y. Guo, X. Huang, and L. Wu. Approaches to event-focused summarization based on named entities and query words. In *Document Understanding Conference (DUC'03)*, 2003.

[GHZ93] Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307, 1993.

[GJW95] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.

[GKMC99] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of ACM SIGIR'99*, pages 121–128, 1999.

[GMCK00] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowiz. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL Workshop on Summarization*, pages 40–48, 2000.

[GMMM00] C. Grover, C. Matheson, A. Mikheev, and M. Moens. Lt ttt: A flexible tokenization toolkit. In *Proceedings of LREC'00*, 2000.

[Gra98] Gregory Grafenstette. Producing intelligent telegraphic text reduction to provide an audio scanning service. In *Intelligent Text Summarization, AAAI Spring Symposium Series*, pages 111–117, 1998.

[Gri75] H. Paul Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Syntax and semantics*, volume 3, pages 43–58. Academic Press, New York, 1975.

[GS86] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204, 1986.

[HO04] Donna Harman and Paul Over. The effects of human variation in duc summarization evaluation. In *Text summarization branches out workshop at ACL'2004*, 2004.

[Hut87] J. Hutchins. Summarization: Some problems and methods. In *Proc. Informatics 9: MeaningThe Frontier of Informatics*, pages 151–173, 1987.

[JBME98] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*, 1998.

[Jin00] Hongyan Jing. Sentence simplification in automatic text summarization. In *Proceedings of the 6th Applied NLP Conference, ANLP'2000*, 2000.

[JM99] Hongyan Jing and Kathleen McKeown. The decomposition of human-written summary sentences. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999.

[JM00] Hongyan Jing and Kathleen McKeown. Cut and paste based text summarization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, 2000.

[Joh70] Ronald Johnson. Recall of prose as a function of structural importance of linguistic units. *Journal of Verbal Learning and Verbal Behaviour*, 9:12–20, 1970.

[KM00] Kevin Knight and Daniel Marcu. Statistics-based summarization — step one: Sentence compression. In *Proceeding of The American Association for Artificial Intelligence Conference (AAAI-2000)*, pages 703–710, 2000.

[KM02] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1), 2002.

[KPC95] Julian Kupiec, Jan Perersen, and Francine Chen. A trainable document summarizer. In *Research and Development in Information Retrieval*, pages 68–73, 1995.

[Kri80] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Sage Publications, Beverly Hills, CA, 1980.

[Kri04] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–437, 2004.

[KT02] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications, 2002.

[LDF05] Jimmy Lin and Dina Demner-Fushman. Evaluating summaries and answers: Two sides of the same coin? In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Measures for MT and Summarization*, 2005.

[Lev84] Elena Levy. *Communicating Thematic Structure: The Use of Referring Terms and Gestures in Narrative Discourse.* PhD thesis, 1984.

[LH00] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, 2000.

[LH02] Chin-Yew Lin and Eduard Hovy. Automated multi-document summarization in neats. In *Proceedings of the Human Language Technology Conference (HLT2002)*, 2002.

[LH03] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*, 2003.

[LHHN04] Finley Lacatusu, Andrew Hickl, Sanda Harabagiu, and Luke Nezda. Lite_gistexter at duc2004. In *Proceedings of the 4th Document Understanding Conference (DUC'04)*, 2004.

[Lin04] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop in Text Summarization, ACL'04*, 2004.

[LO04] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.

[Luh58] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[Mar00] Daniel Marcu. *The Theory and Practice of Discourse and Summarization.* The MIT Press, 2000.

[Mar01] Daniel Marcu. Discourse-based summarization in duc 2001. In *Document Understanding Conference 2001*, 2001.

[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 (demo). In *Proceedings of NAACL-HLT'03*, 2003.

[MBE$^+$01] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Barry Schiffman, and Simone Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of DUC 2001*, 2001.

[MBE$^+$02] Kathleen McKeown, Regina Barzilay, David Evans, Vasleios Hatzivassiloglou, Judith Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the 2nd Human Language Technologies Conference HLT-02*, 2002.

[Mel88] Igor Mel'cuk. *Dependency syntax: theory and practice.* State University of New York Press, 1988.

[MG01] Daniel Marcu and Laurie Gerber. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, 2001.

[MGB99] Inderjeet Mani, Barbara Gates, and Eric Bloedorn. Improving summaries by revising them. In *proceedings of the annual Meeting of the Association for Computational Linguistics (ACL'99)*, 1999.

[MKH$^+$02] Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, and Therese Firmin abd Beth Sundheim. Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68, 2002.

[MNP97] Jean-Luc Minel, Sylvaine Nugier, and Gerald Piat. How to appreciate the quality of automatic text summarization? In *Proceedings of the ACL/ECL'97 Workshop on Intelligent Scalable Text Summarization*, pages 25–30, 1997.

[MPE$^+$05] Kathleen McKeown, Rebecca Passonneau, David Elson, Ani Nenkova, and Julia Hirschberg. Do summaries help? a task-based evaluation of multi-document summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retreival (SIGIR 2005)*, 2005.

[MR95] Kathleen McKeown and Dragomir Radev. Generating summaries of multiple news articles. In *Proceedings of the ACM Conference on Research and Development in Information Retreival (SIGIR'95)*, 1995.

[MT04] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of EMNLP 2004*, pages 404–411, 2004.

[Nen05] Ani Nenkova. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI'05)*, 2005.

[Ng05] Vincent Ng. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Anual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005.

[NP04] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*, 2004.

[NV05] Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. Technical Report MSR-TR-2005-101, Microsoft research, 2005.

[OY04] Paul Over and James Yen. An introduction to duc 2004 intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2004*, 2004.

[Pai90] C. D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing Management*, 26(1):171–186, 1990.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[PN03] Rebecca J. Passonneau and Ani Nenkova. Evaluating content selection in human- or machine-generated summaries: The pyramid method. Technical Report CUCS-025-03, Columbia University, 2003.

[PNMS05] Rebecca Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigleman. Pyramid evaluation ot duc 2005. In *Proceedings of the Document Understanding Conference (DUC'05)*, 2005.

[Pri92] Ellen Prince. The zpg letter: subject, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins, 1992.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, 2002.

[PV98] M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.

[RBGZ01] Dragomir Radev, Sasha Blair-Goldensohn, and Zhu Zhang. Experiments in single and multi-document summarization using MEAD. In *First Document Understanding Conference*, New Orleans, LA, September 2001.

[RBGZSR01] Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Demo Presentation, Human Language Technology Conference*, San Diego, CA, March 2001.

[RJB00] Dragomir Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, 2000.

[RKCZ03] Stefan Riezler, Tracy King, Richard Crouch, and Annie Zaenen. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *HLT-NAACL*, 2003.

[RM97] Dragomir R. Radev and Kathleen R. McKeown. Building a generation knowledge source using internet-accessible newswire. In *Proceedings, Fifth ACL Conference on Applied Natural Language Processing ANLP'97*, pages 221–228, Washington, DC, April 1997.

[RRS61] G. J. Rath, A. Resnick, and R. Savage. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation*, 2(12):139–208, 1961.

[RT03] Dragomir Radev and Daniel Tam. Single-document and multi-document summary evaluation via relative utility. In *Poster session, proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'03)*, 2003.

[RTSL03] Dragomir Radev, Simone Teufel, Horacio Saggion, and W. Lam. Evaluation challenges in large-scale multi-document summarization. In *ACL*, 2003.

[SG01] Natalie Schenker and Jane Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

[SG04] H. Saggion and R. Gaizaukas. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Document Understanding Conference (DUC04)*, 2004.

[Sid02] Advaith Siddharthan. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 60–65, Philadelphia, USA, 2002.

[Sid03] Advaith Siddharthan. *Syntactic simplification and Text Cohesion*. PhD thesis, University of Cambridge, UK., 2003.

[SL05] Caroline Sporleder and Mirella Lapata. Discourse chunking and its application to sentence compression. In *Proceedings of the Human Language Technology Confer-*

ence and the Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'05), 2005.

[SM03] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *HLT-NAACL*, 2003.

[SMC01] B. Schiffman, Inderjeet. Mani, and K. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*, Toulouse, France, July 2001.

[SMG88] A. Sanford, K. Moar, and S. Garrod. Proper names as controllers of discourse focus. *Language and Speech*, 31(1):43–56, 1988.

[SNM02] Barry Schiffman, Ani Nenkova, and Kathleen McKeown. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*, 2002.

[SNM04] Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. Syntactic simplificantion for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 2004.

[SOC$^+$02] Judith Schlesinger, Mary Ellen Okurowski, John Conroy, Dianne O'Leary, Anthony Taylor, Jean Hobbs, and Harold Wilson. Understanding machine performance in the context of human performance for multi-document summarization. In *Proceedings of the 2nd Document Understanding Conference (DUC'02)*, 2002.

[SSJ01] Tetsuya Sakai and Karen Sparck-Jones. Generic summaries for indexing in information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 190–198, 2001.

[Str98] Michael Strube. Never look back: An alternative to centering. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)*, pages 1251–1257, 1998.

[TJ97] Pasi Tapanainen and Timo Jrvinen. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71. Association for Computational Linguistics, 1997.

[TvH04a] Simone Teufel and Hans van Halteren. Agreement in human factoid annotation for summarization evaluation. In *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC'2004)*, 2004.

[TvH04b] Simone Teufel and Hans van Halteren. Evaluating information content by factoid analysis: human annotation and stability. In *EMNLP-04*, 2004.

[VBM04] Lucy Vanderwende, Michele Banko, and Arul Menezes. Event-centric summary generation. In *Proceedings of the Document Understanding Conference (DUC'04)*, 2004.

[vD77] T.A. van Dijk. Semantic macro-structures and knowledge frames in discourse comprehension. *Cognitive Processes in Comprehension*, pages 3–32, 1977.

[vHT03] Hans van Halteren and Simone Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*, 2003.

[Voo04] Ellen M. Voorhees. Overview of the trec 2003 question answering track. In *Proceedings of the 12th Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.

[VS05] Lucy Vanderwende and Hisami Suzuki. Frequency-based summarizer and a language modeling extention. In *MSE 2005 common data task evaluation*, 2005.

[WF05] Ian Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.

[WRC97] N. Wacholder, Y. Ravin, and M. Choi. Disambigaution of names in text. In *Proceedings of the Fifth Conference on Applied NLP*, pages 202–208, 1997.

[ZDL$^+$05] David Zajic, Bonnie Dorr, Jimmy Lin, Christof Monz, and Rich Schwartz. A sentence-trimming approach to multi-document summarization. In *Proceedings of the 5th Document Understanding Conference (DUC'05*, 2005.

[Zip65] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort.* Hafner Publishing Company, 1965.

Appendix A

Summary content units (SCUs)

The goal of SCU annotation is to identify sub-sentential content units that can allow for comparison of the information in several summaries. It is well-known that when summarizing people make different choices about what information to include in their summary. The SCU annotation aims at highlighting what people agreed on. After the annotation is completed, some SCUs might appear in only one summary, but its annotation will allow a person to read a brand new summary and look for that SCU in this new summary.

An SCU consist of a label and contributors. The label is a concise English sentence that states the semantic meaning of the content unit. The contributors are snippet(s) of text coming from the summaries that show the wording used in a specific summary to express the label. It is possible for an SCU to have a single contributor, in the case when only one of the analyzed summaries expresses the label of the SCU.

The definition of content unit is somewhat fluid, it can sometimes be a single word but it is never bigger than a sentence clause. Any event realized by a verb or a nominalized verb (e.g, "blow up" and "bombing" in the examples below) is a candidate SCU.¡/P¿

The three questions that will help you identify an SCU contributor are

1. Is the information expressed by it repeated in some other summary? Note, the wording need not be the same for the expressed meaning to be the same; we are looking for the same meaning. When an information unit is expressed in two or more summaries, the amount of information overlap will serve as a main indication of which parts of

the corresponding sentences will become contributors.

2. Spans of words that indicate location or time, or otherwise provide more specific information about another SCU are also SCUs. Usually these are expressed in adjuncts such as prepositional phrases and are not an obligatory argument to any verb. Noun phrases containing premodification can also be split into more than one SCU when the premodifiers include additional information. The need to split such additional information will arise in two cases. 1) When more than one summary express some information, but one of the summaries has an adjunct, e.g. several summaries mention that there was a bombing and one summary mentions the exact location of the bombing. In this situation one would identify two SCUs, one with the main event, and one with the additional detail information.

3. Is the difference important for the story? Occasionally there will be minor differences in wording that if put under scrutiny could be construed to have different nuances. We are not interested in the finest grained distinctions—these will be too many to describe in a reasonable way.

Example 1: The three sentences below come from four different summaries A, B, C and D.

A: In 1992 the U. N. voted sanctions against Libya for its refusal to turn over the suspects.

B: The United Nations imposed sanctions on Libya in 1992 because of their refusal to surrender the suspects.

C: The U.N. imposed international air travel sanctions on Libya to force their extradition.

D: Since 1992 Libya has been under U.N. sanctions in effect until the suspects are turned over to United States or Britain.

Among other information, all four sentences express the fact that "Libya was under U.N. sanctions" and this is the label for the SCU. The contributors are marked in brackets below (ignore SCU2 for now.)

A: In 1992 [the U. N. voted sanctions against Libya]1 [for its refusal to turn over the suspects.]2

B: [The United Nations imposed sanctions on Libya]1 in 1992 [because of their refusal to surrender the suspects.]2

C: [The U.N. imposed]1 international air travel sanctions on Libya [to force their extradition.]2

D: Since 1992 [Libya has been under U.N. sanctions]1 [in effect until the suspects are turned over]2 to United States or Britain.

Other information, such as when the sanctions where imposed, what specific sanctions were imposed, why they were imposed etc, will form their own SCUs. Identifying a main topic event in the summaries and asking yourself such questions as above about specifics will help you formulate labels and identify the SCU contributors. The contributors of an SCU need not share identical wording. For example in the sentences above, the SCU with label "The goal behind the sanctions is to make Libya surrender the suspects" is expressed by the text coindexed with "2". Sentence B differs in wording from the rest of the sentences, but the meaning is the same as that of the other contributors, expressing the fact that Libya does not want to surrender the suspects and the other nations involved want to force their extradition. (Note that this is an example of only two SCUs that will be derived from the sentences, the full analysis will lead to identifying more SCUs and will lead to complete bracketing of the sentences.)

Let's look at one more example of sentences from the different summaries that share some common information.

A. In 1998 [two Libyans indicted]1 [in 1991]2 for the Lockerbie [bombing]3 were still in Libya.

B. [Two Libyans were indicted]1 [in 1991]2 [for blowing up]3 [a Pan Am]5 [jumbo jet]4 over Lockerbie, Scotland in 1988.

C. [Two Libyans, accused]1 by the United States and Britain [of bombing]3 [a New York bound]6 [Pan Am]5 [jet]4 over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.

D. [Two Libyan suspects were indicted]1 [in 1991]2.

All share the information that (1) "Two Libyans are held responsible for a crime". The contributors are surrounded by brackets and coindexed by 1. Note that C differs in its wording from the other sentences–accused is not the same as indicted. But because the goal of the annotation is to find as much shared information as possible, and the sense of "accused" is so close to that of "indicted", the contributors will be grouped together, and the label expresses the general meaning of both accused and indicted.

The time expression prepositional phrase "in 1991" forms a separate SCU because the phrase "in 1991" can be omitted for example from sentence D without making the sentence ungrammatical or incomprehensible. There will be loss of information, and this is why the phrase can indicate a new *content* unit! The contributors of the SCU with label "The libyans were accused in 1991" are coindexed with "2".

Now we have to proceed and find what other information is repeated. For example, what was the crime committed? The different sentences give different amount of detail. When deciding where to start from–remember that the main goal is identifying the same information! All sentences agree on the fact that "the crime in question is a bombing" – the contributors are coindexed with 3.

What was bombed? "An airplane was bombed" is another SCU with index 4. This information is expressed in two bigger noun phrases " Pan Am jumbo jet" and "a New York bound Pan Am jet" but "New York bound" and "Pan Am" can be omitted and the sentences will still be acceptable, so this information will be marked in a separate content unit.

The contributors are simply a part of the sentence–not all grammatical arguments necessary to reconstruct the label will be included in the contributor. This is ok, because the label will "bring in" any argument needed.

It is best if the SCU contributor can be a complete grammatical phrase. But this is sometimes not possible, so use your best judgment in assigning the specific token boundaries of the contributor.

Appendix B

Seven translations of the same text

B.1 ta0/chtb_003.sgm

Fourteen Chinese Open Border Cities Make Significant Achievements in Economic Construction

Xinhua News Agency report of February 12 from Beijing - The fourteen Chinese border cities that have been opened to foreigners achieved satisfactory results in their economic construction in 1995. According to statistics, the cities achieved a combined gross domestic product of RMB19 billion last year, an increase of more than 90% over 1991 before their opening. The State Council successively approved the opening of fourteen border cities to foreigners in 1992, including Heihe, Pingxiang, Hunchun, Yining and Ruili, and permitted them to set up 14 border economic cooperation zones. Over the past three years, the cities have undergone rapid social and economic development, considerably enhancing their local economic strength. They achieved an average annual economic growth of 17%, higher than the national average. According to the relevant sources, the 14 cities have stepped up their urban construction and the development and construction of the cooperation zones. During the past three years, these border cities maded a combined investment of RMB12 billion in fixed assets and have changed the past situation of ”no high buildings, no smooth roads, no bright lights, no clear water and no easy communication”. A total of 22.6 square kilometers of economic cooperation zones have been developed, attracting a total of 287 foreign-funded enterprises with an actual foreign investment of US$890 million. In addition, there are some

5,100 domestic enterprises, with 175 industrial projects already put into operation.

B.2 ta1/chtb_003.sgm

Significant Accomplishment Achieved in the Economic Construction of the Fourteen Open Border Cities in China

Xinhua News Agency, Beijing, Feb.12 - Exciting accomplishment has been achieved in 1995 in the economic construction of China's fourteen border cities open to foreigners. Statistics have indicated that these cities produced a combined GDP of over 19 billion yuan last year, an increase of more than 90%, compared with that in 1991 before the cities were open to foreigners. In 1992, the State Council successively opened fourteen border cities to foreigners. These included Heihe, Pingxiang, Huichun, Yining, and Ruili. Meanwhile, the State Council also gave its approval to these cities to establish fourteen border zones for economic cooperation. The past three years saw a rapid social and economic development in these cities; the local economic power enjoyed a significant boost; and the annual economic growth rate has averaged 17%, exceeding that of the national average. It is reported that the urban construction in these fourteen cities and the development of the cooperation zones are speeding up. Over the past three years, these cities have invested 12 billion yuan in fixed assets. The old image of the border cities invoking " low buildings, uneven roads, dim lights, muddy water and poor communication" has changed. Within the economic cooperation zones, a total of 22.6 square kilometers of land has been developed; 287 "Three-Capital" ventures have been invited to move in with actual utilization of foreign capital of 890 million US dollars. In addition, there are 5,100 inland associated enterprises with 175 industrial projects already in operation.

B.3 ta2/chtb_003.sgm

In China, fourteen cities along the border opened to foreigners achieved remarkable economic development

Xinhua News Agency, Beijing, February 12 - The economic development in China's fourteen cities along the border opened to foreigners achieved gratifying results in 1995.

According to statistics, these cities completed a gross domestic product in excess of RMB 19 billion in last year, an increase of more than 90% over 1991 (the year before they were opened). In 1992, the State Council successively approved fourteen cities along the border to be opened to foreigners, which included Hei He, Pingxiang, Hunchun, Yining and Ruili etc. At the same time, these cities were also given approvals to set up fourteen border-economic-cooperation zones. The economy in these cities grew at a rapid rate in the past three years and their economic strength are now noticeable stronger than before; the average annual economic growth is 17%, which is higher than the average growth rate in China. According to introduction, developments in these fourteen cities as well as their economic cooperation zones are still accelerating. In the past three years, these cities completed fixed assets investments to the value of RMB 12 billion, which has changed the typical scenario prevailing in border cities in the old days, namely, "no high rise building, uneven roads, no bright lights, murky water and poor telecommunication". The economic cooperation zones completed a development of 22.6 square kilometers; 287 firms of the "three forms" of capitals were introduced and the actual foreign capital utilized was US$ 890 million. Furthermore, there are also 5,100 domestic linked enterprises and 175 industrial projects currently in production.

B.4 ta3/chtb_003.sgm

Economic Construction Achievement is Prominent in China's Fourteen Border Opening-up Cities.

Xinhua News Agency, Beijing, February 12 - delightful economic construction result was achieved in China's fourteen border opening-up cities in 1995. According to statistics, GDP registered over 19 billion yuan last year in those cities, over 90% higher than those of year 1991 before opening-up. Fourteen border cities like Heihe, Pingxiang, Huichun, Yinin, and Ruili etc were approved successively by the State Council in 1992 as the cities opening to the outside world, setting up of fourteen border economic co-operation zones in these cities were also approved simultaneously. Since more than three years, the social economy has been developed quickly in these cities with considerably strengthened local economic power; their average annual economic growth is 17% that is higher than the average annual growth rate nationwide. It is said that the urban construction and the construction and development of co-operation zone in these 14 cities have been speeded up. Since three years, the accumulative fixed investment in these cities has reached 12 billion yuan, the circumstances of border cities in former days, "buildings were not tall, roads were not level, lamps were not bright, water was not clean, communication was not expedite," has been changed. In the economic co-operation zone, 22.6 square kilometers of land was developed, 287 foreign-founded enterprises were introduced, with an actual utilization of foreign investment at 890 million US dollars. In addition, there are 5,100 domestic-cooperative enterprises, and 175 industrial projects that were put into production, in these zones.

B.5 ta4/chtb_003.sgm

Fourteen Chinese frontier cities see significant economic achievement

Xinhua News Agency, Beijing, February 12. Fourteen frontier cities that adopted the open-door policy realized significant results in their economic growth in 1995. Statistics show that the total GDP of these cities last year was more than 19 billion RMB, an increase of more than 90% over 1991 when the cities had not yet opened. The State Council in 1992 approved fourteen frontier cities as cities opened to the outside. These cities included Heihe, Pingxiang, Huichun, Yining, and Ruili. The State Council also approved the establishment of 14 economic cooperation zones in these cities. In the past three years, these cities have undergone rapid social economic development, with the strength of the local economies increasing visibly. Anual growth rate in these cities has averaged 17%, greater than the average national rate of growth. These 14 cities expedited their pace of urban construction and development of their cooperation zones. In the past three years, these cities accumulated fixed asset investment of RMB12 billion. The low buildings, rugged roads, dim lights, dirty water, and poor communication infrastructure of the past have all been improved. Am area of 22.6 square kilometers has been developed in the economic cooperation zones. The areas have attracted 287 WOFEs (wholly owned by foreign investment enterprises) and JVs (joint ventures), with actual foreign investment of US$890,000,000. In addition, these cities also have some 5,100 domestic businesses and 175 industrial projects have been in operation.

B.6 ta5/chtb_003.sgm

Significant Economic Construction Results in Fourteen Open Chinese Border Cities

Xinhua News Agency, February 12, Telegram. Fourteen Chinese border cities open to foreign investment have achieved encouraging economic construction results in 1995. According to statistics, these cities achieved domestic production of more than 19 billion yuan in total value last year, amounting to growth of more than 90 percent over 1991, the year before these cities were opened up. The State Council successively approved the opening up of fourteen border cities including Heihe, Pingxiang, Hunchun, Yining, and Ruili in 1992, simultaneously approving the establishment of fourteen economic cooperation border zones by these cities. Over the past more than three years, the socioeconomic development of these cities has been rapid and local economic strength has increased markedly; the economy has grown at an average annual rate of 17 percent, which is higher than the average growth rate of the nation as a whole. Based on information provided, the pace of municipal construction and cooperation zone development construction in these fourteen cities is accelerating. Over the past three years, these cities have accumulated 12 billion yuan in fixed capital investment, changing the old image of border cities having "squat buildings, rough roads, lack of lighting, unclean water, and poor communications". Within the economic cooperation zones, 22.6 square kilometers have been developed, 287 "three-capital" enterprises have been brought in, and US$890 million in foreign capital has actually been utilized. Additionally, 5,100 domestic affiliate enterprises have put 175 industrial projects into operation.

B.7 ta6/chtb_003.sgm

China's 14 Border Open Cities Achieved Sound Economic Development

Beijing, February 12 (Xinhua News Agency) China's 14 border open cities achieved gratifying economic progress in 1995. As statistics show, they achieved a GDP growth above 90% of what they did before they had not been border open cities in 1991. In 1992, the State Council successively ratified 14 cities by the border, namely Heihe, Pingxiang, Huichun, Yining, Ruili, etc, as open-up cities, allowing them to set up 14 border economic cooperation zones. Three years later, all these cities rapidly developed their social economy and distinctly strengthened their locale economic capacity, achieving a yearly economy increase of 17%, which is higher than the state average standard. As it is introduced, in these 14 cities, the construction of urban area and open-up zones are becoming more rapid. 3 years later, these cities have fulfilled RMB 12 billion of fix assets investment and have changed the old day's situation in border cities, such as low buildings, bumpy roads, dim lights, muddy water and blocked communications. The economic cooperation zones have opened up an area of 22.6 square km, invited in 287 joint ventures and made use of exactly 890 million US dollars of foreign capital. Furthermore, there are 5100 internal cooperative ventures and 75 industrial projects already carried out.

Appendix C

Seven summaries produced for the same input texts

C.1 D311.M.250.I.D

The industrial espionage case involving GM and VW began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as production director. Ferdinand Piech had just been installed as Chairman of Volkswagen when he hired Lopez. Lopez brought with him to VW seven former GM employees including his close friend, Jorge Manuel Gutierrez. When GM protested the recruiting of their employees, private mediation was set up between the two companies. The situation became more serious when top- secret documents were found missing from GM that described plans to build a new model car. Then state prosecution officials discovered four boxes of papers, including secrets documents containing plans to build the car where Gutierrez and Piazza, another former GM employee were staying. This led to investigations and counter charges on both sides of the ocean followed by civil and criminal court cases. Lopez was found innocent during his first trial in Germany, but faced further charges. A court case was also brought against leading news magazine, Der Spiegel, for making public allegations of spying against Lopez. Der Spiegel later presented evidence in state court in a bid not to be stopped from reporting. The case reached the highest levels in both countries including the US Justice Dept. and the FBI investigating mail and wire fraud. The legal cases soon became bogged down in mountains of papers and transatlantic issues between the countries involving lawyers and politicians. As of October 1994, nothing had been resolved.

C.2 D311.M.250.I.E

Jose Lopez as head of purchasing at Adam Opel, GMs German subsidiary, led the company to become the most profitable car maker in that country. He was made procurement chief at GM headquarters. However, he left GM for VW under circumstances, which along with ensuing events, were described by a German judge as "potentially the biggest-ever case of industrial espionage". Ferdinand Piech took over an ailing VW company that was losing money. He is accused of luring Lopez away from GM along with several of Lopez' key associates. Lopez, disappointed that GM was not going to build a plant in his native Basque country, left GM under a cloud of confusion and allegedly took some sensitive GM documents and plans with him. Lopez also requested documents from Adam Opel that later turned up in the Wiesbaden home of a Lopez colleague who followed him to VW. Still Later German police raided VW headquarters and seized documents that were later founds to contain Opel plans. Months of charges and counter-charges followed and details were aired in German newspapers. German economics minister Gunter Rexrodt was concerned of damage to US-German political and business relations and for a time tried to be a peacemaker. PM Gerhardt Schroeder of Lower Saxony, VW's largest shareholder, strongly supported VW. Lopez agreed to pay DM75,000 instead of facing perjury charges in court. Criminal investigations in Germany and the US relating to suspected industrial espionage, theft, perjury and wire fraud are still in progress.

C.3 D311.M.250.I.F

The industrial espionage case involving Volkswagen and General Motors began when VW failed to convince GM that its plans for a revolutionary automobile plant in Spain were not copies of a proposed GM project. As early as December, 1992, Jose Ignacio Lopez de Arriortua, of GM, was in touch with Ferdinand Piech, Chairman of Volkeswagen, about coming to work for VW. Lopez was procurement chief at Adam Opel, GM's German subsidiary. Lopez was disappointed by GM's decision not to build an automobile plant in his own Basque country. VW offered to build it. He left GM for VW in March, 1993. He also took, according to considerable evidence, many GM documents, along with seven GM executives. In April, 1993, witnesses in Wiesbaden allegedly saw documents being shredded by Jorge Alvarez Aquirre and Rosario Piazza, two Lopez associates. In four remaining boxes, investigators found details of Opel secret car plans. A VW employee said she had punched Opel data into the VW computer. In July, the U.S. Justice Department announced it was investigating the Lopez case. At a VW meeting in August, 1993, Lopez contradicted his earlier public claim that he never took any secret documents and said that papers from his former offices were destroyed in order to keep them from being circulated within VW. Investigations continued. In May, 1994, Lopez agreed to pay 29,850 pounds to avoid facing perjury charges in court. Through October, 1994, no legal action had been taken against Lopez or Volkswagen.

C.4 D311.M.250.I.G

The industrial espionage battle by General Motors (GM) and its German subsidiary, Adam Opel, against Volkswagen (VW) for the past 18 months has bogged down in mountains of paper and a complex transatlantic tussle involving both lawyers and politicians. Germany, increasingly concerned about the effect of the court investigations on domestic economic and political affairs, is now also worried about Bonn's relations with Washington. A US probe of the investigation started at the instigation of the Commerce Department after President Clinton apparently decided that industrial espionage in general was a threat to America's well being. The FBI is probing possible mail and wire fraud. The issue stems from the alleged recruitment of GM's eccentric and visionary Basque-born procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues. He was presumably recruited by VW chairman, Ferdinand Piech, who was impressed with Lopez's leading role in helping Adam Opel recover from a major production cost disadvantage. He saw Lopez as the answer to a similar ongoing problem at VW. The investigation is focused mainly on evidence that Mr. Lopez and his associates took GM and Adam Opel industrial secrets with them. These included details of Opel's entire European component supplier network and key contact data, plans for a new style, low-cost high-speed car factory, and information on new models. Coincidentally, Lopez quit after being informed that a plan to install his new car dream plant in his Basque area was cancelled. The outcome of the investigation is still uncertain.

C.5 D311.M.250.I.H

On March 16, 1993, with Japanese car import quotas to Europe expiring in two years, renowned cost-cutter, Agnacio Lopes De Arriortua, left his job as head of purchasing at General Motor's Opel,Germany, to become Volkswagen's Purchasing and Production Director. GM charged that during his last months at GM, Lopez stole GM plans for ultra-low cost factories; designs for advanced cars and engines; and information about Opel's suppliers and parts. A regional court in Frankfurt issued an injunction preventing VW from recruiting more GM staff. All charges of anti-competitive staff poaching were later dismissed. VW failed to get court injunctions preventing Der Spiegel magazine from publishing GM's allegations. Darmstadt, Germany investigator, Dorothea Holland, launched still unresolved investigations into charges of industrial espionage, theft, and perjury against Lopez, his 22-year old daughter, Begounia, and other GM colleagues now at VW. VW counter-charged that GM had planted GM documents and data in VW sites and computers. The FBI opened still unresolved investigations of wire and mail fraud against VW and Lopez. European government and industry leaders expressed fear that defensive allegations by VW Chairman Ferdinand Piech, that the Lopez incident amounted to U.S. industrial warfare against Germany, would destabilize U.S.-European commercial and diplomatic relations. U.S. Commerce Secretary, Ron Brown, suggested that relations between the U.S. and Germany would be damaged if German investigators don't immediately deliver long promised data and assistance in the GM/VW case. So far, Lopez has agreed to pay DM75,000 to set aside the perjury case against him.

C.6 D311.M.250.I.I

Ferdinand Piech became Volkswagen chairman in January 1993 and planned to turn the money-losing company around. A ruthless restructurer, he recruited General Motors' like-minded global head of purchasing Jose Ignacio Lopez de Arriortua. Lopez had developed a new GM plant to be built in his native Basque country, but when he learned in March 1993 that it would be built in Hungary he abruptly left GM for VW, with secret GM documents he had requested that detailed GM's new plant and car models. Lopez was followed by seven top members of his team, including Jose Gutierrez, Jorge Alvarez Aguirre, and Rosario Piazza. He tried to recruit others. Soon after Lopez's arrival, VW announced a new plant to be built in Basque country and a car similar to a planned GM model. Piech publicly accused GM/Opel of planting documents and hacking VW computers with the aim of destroying VW. This angered GM and led Germany to distance itself from Piech's damaging nationalistic tones. Piech's clumsy, halfhearted conciliation efforts failed. At GM's request, German state prosecutors and the FBI began investigations into industrial espionage, theft, perjury and wire fraud. GM wanted all former employees banned from working for VW for 12 months but settled for only some. VW was banned from further personnel poaching but a Frankfurt court denied that poaching broke fair competition rules. Lopez paid a fine instead of facing perjury charges. German politicians called the case biased and wanted it dropped. In October 1994 criminal charges were bogged down.

C.7 D311.M.250.I.J

General Motors Corporation and Volkswagen were warring in 1993 and 1994. VW chairman Ferdinand Piech recruited the General Motors/Opel executive, Jose Lopez de Arriortua. In March 1993, Lopez and seven other GM executives moved to VW overnight. GM immediately accused Lopez of looting Opel's supply network and contract database and taking secret plans for a high-speed factory and a new Opel mini-car. VW and Lopez also were accused on conducting an illegal recruiting campaign. German officials began investigating VW for theft and industrial espionage. With GM urging, a temporary injunction was imposed on VW recruiting, but it was subsequently lifted and manager-poaching claims against VW were rejected. GM documents and computerized information were seized from a VW headquarters and documents were found at the apartment of the former GM executives, Jorge Alvarez Aquirre and Rosario Piazza. The German economics minister, Gunter Rexrodt, had tried to be a peacemaker is this controversy, but in September 1993 withdrew. Lopez was accused of perjury and in May 1994 agreed, while maintaining his innocence, to pay a DM75,000 fine to avoid facing charges in court The German prosecutor, Dorthea Holland, was searching through an estimated 2 million computer printout sheets. Because of leaks, a gag was placed on her office in October 1994 and no information was expected until a decision to indict Lopez was reached. The U.S. Justice Department's interest in industrial espionage had been piqued and the FBI began an investigation of mail and wire fraud, which was also stalled.