Text Summarization

Today

- HW3 due
- Summarization
  - Introduction
  - Extractive methods
  - Abstractive methods and Columbia’s newsblaster
  - Evaluation
  - Updates on disaster

What is Summarization?

- Data as input (database, software trace, expert system), text summary as output
- Text as input (one or more articles), paragraph summary as output
- Multimedia in input or output
- Summaries must convey maximal information in minimal space

Why is Summarization Hard?

- Seems to require both interpretation and generation of text
- Handle input documents from unrestricted domains robustly
- Operate without full semantic interpretation

Leads many summarization researchers to use sentence selection

Types of Summaries

- Informative vs. Indicative
  - Replacing a document vs. describing the contents of a document
- Extractive vs. Generative (abstractive)
  - Choosing bits of the source vs. generating something new
- Single document vs. Multi Document
- Generic vs. user-focused

Types of Summaries

- Informative vs. Indicative
  - Replacing a document vs. describing the contents of a document
- Extractive vs. Generative (abstractive)
  - Choosing bits of the source vs. generating something new
- Single document vs. Multi Document
- Generic vs. user-focused

Questions (from Sparck Jones)

- Should we take the reader into account and how?

- “Similarly, the notion of a basic summary, i.e., one reflective of the source, makes hidden fact assumptions, for example that the subject knowledge of the output’s readers will be on a par with that of the readers for whom the source was intended. (p. 5)”

- Is the state of the art sufficiently mature to allow summarization from intermediate representations and still allow robust processing of domain independent material?

Foundations of Summarization – Luhn; Edmunson

- Text as input
- Single document
- Content selection
- Methods
  - Sentence selection
  - Criteria

Sentence extraction

- Sparck Jones:

- `what you see is what you get’, some of what is on view in the source text is transferred to constitute the summary

Luhn 58

- Summarization as sentence extraction
- Term frequency determines sentence importance
    - Stop word filtering
    - Similar words count as one
    - Cluster of frequent words indicates a good sentence

*TF*IDF*

- Intuition: Important terms are those that are frequent in this document but not frequent across all documents

Term Weights

- Local weights
  - Generally, some function of the frequency of terms in documents is used
- Global weights
  - The standard technique is known as inverse document frequency

$$idf_i = \log\left(\frac{N}{n_i}\right)$$

N= number of documents; ni = number of documents with term i

TFxIDF Weighting

- To get the weight for a term in a document, multiply the term’s frequency derived weight by its inverse document frequency.

  TF*IDF

Edmunson 69

Sentence extraction using 4 weighted features:

- Cue words (“In this paper..”, “The worst thing was ..”)
- Title and heading words
- Sentence location
- Frequent key words

Sentence extraction variants

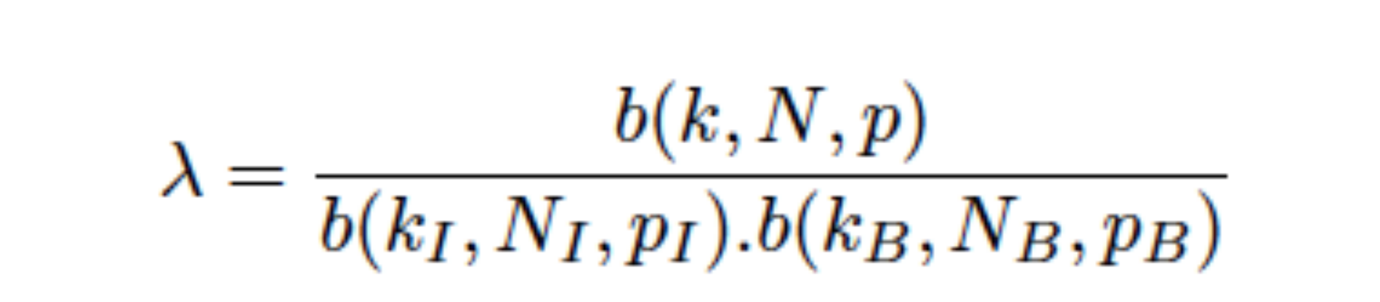
Which sentences are salient in a single document or a document cluster?

- Word frequency variants: Log Likelihood
  - Lin and Hovy
  - Conroy
- Graph based models: Lexrank
  - Erkun and Radev

Topic Signature Words

- Uses the log ratio test to find words that are highly descriptive of the input
- the log-likelihood ratio test provides a way of setting a threshold to divide all words in the input into either descriptive or not
  - the probability of a word in the input is the same as in the background
  - the word has a different, higher probability, in the input than in the background
- Binomial distribution used to compute the ratio of the two likelihoods
- The sentences containing the highest proportion of topic signatures are extracted.

Log likelihood ration



$$\lambda = \frac{b(k, N, p)}{b(k_I, N_I, p_I).b(k_B, N_B, p_B)}$$

Where the counts with subscript i occur in the input corpus and those with subscript B occur in the background corpus

Probability (p) of w occuring k times in N Bernoulli trials

The statistic -2λ has a known statistical distribution: chi-squred

Graph-based methods

- Sentence similarity is measured as a function of word overlap
  - Frequently occurring words link many sentences
  - Similar sentences give support for each other's importance
- Input represented as highly connectived graph
  - Vertices represent sentences
  - Edges between sentences weighted by similarity between two sentences
  - Cosine similarity with TF*IDF weights for words

Sentence Selection

- Vertex importance (centrality) computed using graph algorithms
  - Edge weights normalized to form probability distribution -> Markov chain
  - Compute probablity of of being in each vertex of graph at time t while making consecutive transitions from one vertex to next
  - As more transitions made, probability of each vertex converges -> stationary distribution
- Vertices with higher probability = more important sentences

Example summaries Baseline: first paragraph of last article in cluster

- The Swiss government has ordered no investigation of possible bank accounts belonging to former Chilean dictator Augusto Pinochet, a spokesman said Wednesday. Weekend newspaper reports in Spain said a Spanish judge who ordered Pinochet's arrest has issued a petition aimed at freezing any accounts the 82-year-old general might have in Luxembourg and Switzerland. But government spokesman Achille Casanova said no accounts have so far been frozen in Switzerland and no investigation order has been given to federal banking authorities. Pinochet has been held at a London clinic since his arrest earlier this month.

Example Summaries Topic signatures

- As his lawyers in London tried to quash a Spanish arrest warrant for Gen. Augusto Pinochet, the former Chilean dictator, efforts began in Geneva and Paris to have him extradited. Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker saying that Chile's claim that the former Chilean dictator has diplomatic immunity is ridiculous. Margaret Thatcher entertained former Chilean dictator Gen. Augusto Pinochet at her home two weeks before he was arrested in his bed in a London hospital, the ex-prime minister's office said Tuesday, amid growing diplomatic and domestic controversy over the move.

Example Summaries Lexrank

Cuban President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean dictator Augusto Pinochet calling it a case of international meddling. Pinochet, 82, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. The Chilean government has protested Pinochet's arrest insisting that as a senator he was traveling on a diplomatic passport and had immunity from arrest. Castro, Latin America's only remaining authoritarian leader, said he lacked details on the case against Pinochet but said he thought it placed the government of Chile and President Eduardo Frei in an uncomfortable position.

problems did you see in the automatic summ

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Human Summary

- Former Chilean dictator Augusto Pinochet has been arrested in London at the request of the Spanish government. Pinochet, in London for back surgery, was arrested in his hospital room. Spain is seeking extradition of Pinochet from London to Spain to face charges of murder in the deaths of Spanish citizens in Chile under Pinochet's rule in the 1970s and 80s. The arrest raised confusion in the international community as the legality of the move is debated. Pinochet supporters say that Pinochet's arrest is illegal, claiming he has diplomatic immunity. The final outcome of the extradition request lies with the Spanish courts.

Summarization as a Noisy Channel Model

- Summary/text pairs
- Machine learning model
- Identify which features help most

Julian Kupiec SIGIR 95 Paper Abstract

- To summarize is to reduce in complexity, and hence in length while retaining some of the essential qualities of the original.
- This paper focusses on document extracts, a particular kind of computed document summary.
- Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document, which suggests that even shorter extracts may be useful indicative summaries.
- The trends in our results are in agreement with those of Edmundson who used a subjectively weighted combination of features as opposed to training the feature weights with a corpus.
- We have developed a trainable summarization program that is grounded in a sound statistical framework.

Statistical Classification Framework

- A training set of documents with hand-selected abstracts
  - Engineering Information Co provides technical article abstracts
  - 188 document/summary pairs
  - 21 journal articles
- Bayesian classifier estimates probability of a given sentence appearing in abstract
  - Direct matches (79%)
  - Direct Joins (3%)
  - Incomplete matches (4%)
  - Incomplete joins (5%)
- New extracts generated by ranking document sentences according to this probability

Features

- Sentence length cutoff
- Fixed phrase feature (26 indicator phrases)
- Paragraph feature
  - First 10 paragraphs and last 5
  - Is sentence paragraph-initial, paragraph-final, paragraph medial
- Thematic word feature
  - Most frequent content words in document
- Upper case Word Feature
  - Proper names are important

Evaluation

- Precision and recall
- Strict match has 83% upper bound
  - Trained summarizer: 35% correct
- Limit to the fraction of matchable sentences
  - Trained summarizer: 42% correct
- Best feature combination
  - Paragraph, fixed phrase, sentence length
  - Thematic and Uppercase Word give slight decrease in performance

Questions (from Sparck Jones)

- Should we take the reader into account and how?

- “Similarly, the notion of a basic summary, i.e., one reflective of the source, makes hidden fact assumptions, for example that the subject knowledge of the output’s readers will be on a par with that of the readers for whom the source was intended. (p. 5)”

- Is the state of the art sufficiently mature to allow summarization from intermediate representations and still allow robust processing of domain independent material?

Text *Summarization at Columbia*

- Shallow analysis instead of information extraction
- Extraction of *phrases* rather than sentences
- Generation from surface representations in place of semantics

Problems with Sentence Extraction

- Extraneous phrases
  - “The five were apprehended along Interstate 95, *heading south in vehicles containing an array of gear including … ...* authorities said.”
- Dangling noun phrases and pronouns
  - “The five”
- Misleading
  - ➢Why would the media use this specific word (fundamentalists), so often with relation to Muslims? *Most of them are radical Baptists, Lutheran and Presbyterian groups.

Cut and Paste in Professional Summarization

- Humans also reuse the input text to produce summaries
- But they *"cut and paste"* the input rather than simply extract
  - our automatic corpus analysis
    - 300 summaries, 1,642 sentences
    - 81% sentences were constructed by cutting and pasting
  - linguistic studies

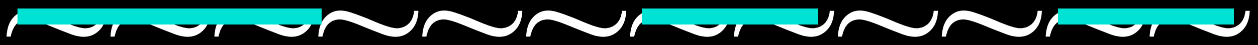
Major Cut and Paste Operations

- (1) Sentence compression

∼∼∼∼∼∼∼∼∼∼∼∼

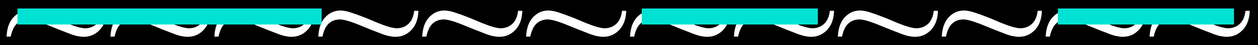
Major Cut and Paste Operations

- (1) Sentence compression

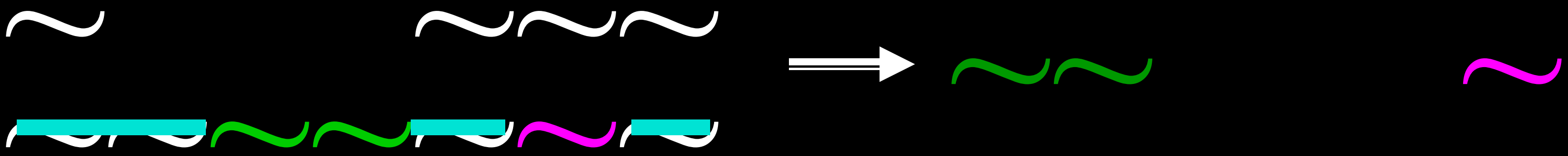


Major Cut and Paste Operations

- (1) Sentence compression

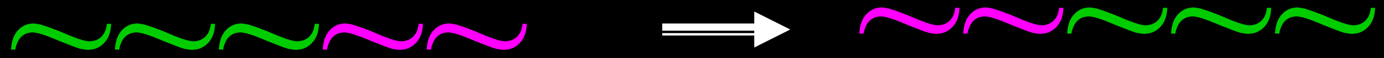


- (2) Sentence fusion

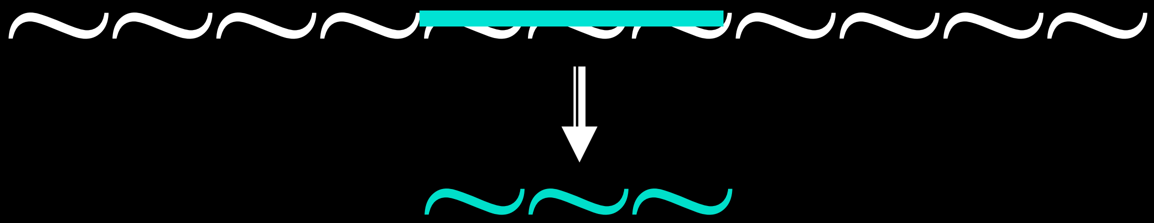


Major Cut and Paste Operations

- (3) Syntactic Transformation



- (4) Lexical paraphrasing



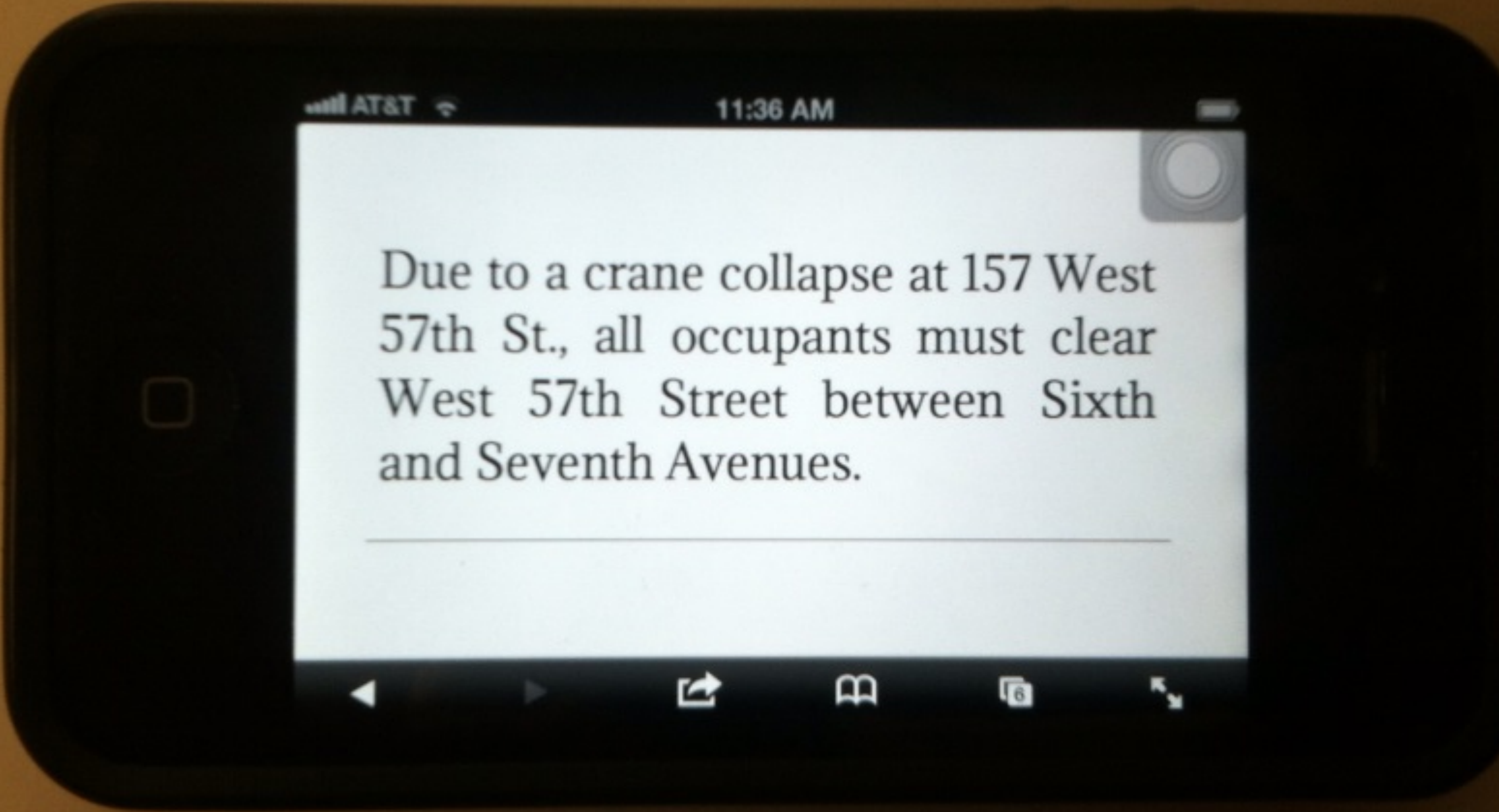
Summarization at Columbia

- News
- Email
- Meetings
- Journal articles
- Open-ended question-answering
    - What is a Loya Jurga?
    - Who is Mohammed Naeem Noor Khan?
    - What do people think of welfare reform?

Summarization at Columbia

- News
- Email
- Meetings
- Journal articles
- Open-ended question-answering
    - What is a Loya Jurga?
    - Who is Mohammed Naeem Noor Khan?
    - What do people think of welfare reform?

Text Compression



Due to a crane collapse at 157 West 57th St., all occupants must clear West 57th Street between Sixth and Seventh Avenues.

Dataset for compression (~3000 sentence pairs)

Clarke & Lapata (2008)

Input

- Italian air force fighters scrambled to intercept a Libyan airliner flying towards Europe yesterday as the United Nations imposed sanctions on Libya for the first time in Col Muammar Gaddafi 's turbulent 22 years in power .

Compression

- Italian air force fighters scrambled to intercept a Libyan airliner as the United Nations imposed sanctions on Libya .

Text to Text Generation

Model text transformation as a *structured prediction* problem

- Input: One or more sentences with parses
- Output: Single sentence + parse

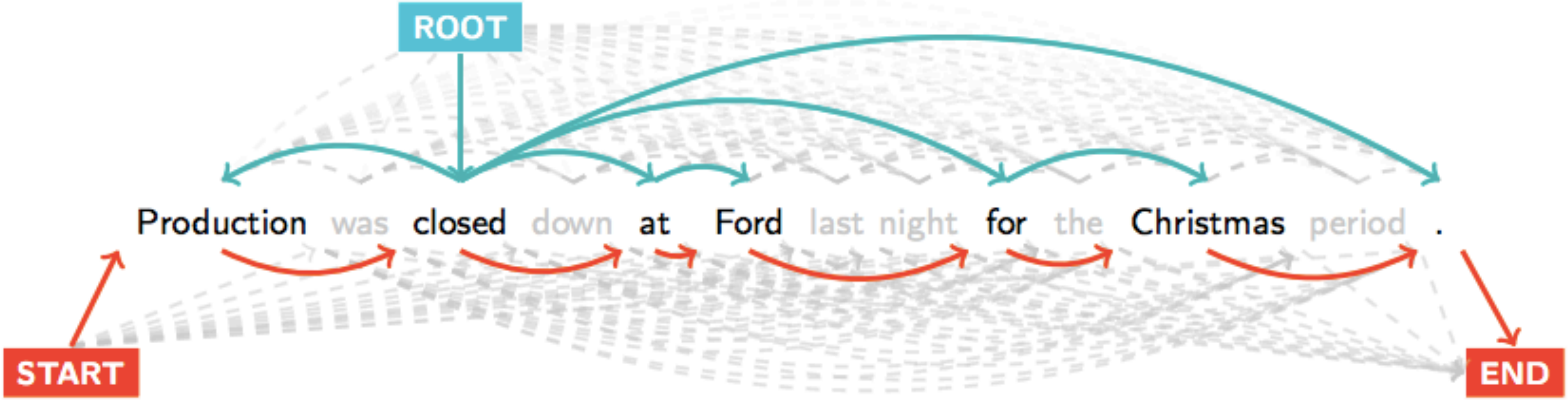
Joint inference over

- **word choice**,
- **n-gram ordering**
- **dependency structure**

Thadani & McKeown, CONLL 2013

structural factorizations

this work



Goal: recover tokens **x**, n-gram sequence **y** and dependency structure **z**

Slide from Thadani

joint inference via ILP

objective

$$
\begin{aligned}
C = \arg\max_{\mathbf{x},\mathbf{y},\mathbf{z}} \; & \boxed{\sum_i x_i \cdot \mathbf{w}_{tok}^{\top} \phi(t_i)} && \text{token score} \\
+ \; & \boxed{\sum_{i,j,k} y_{ijk} \cdot \mathbf{w}_{ngr}^{\top} \phi(\langle t_i, t_j, t_k \rangle)} && \text{ngram score} \\
+ \; & \boxed{\sum_{i,j} z_{ij} \cdot \mathbf{w}_{dep}^{\top} \phi(\langle t_i, t_j \rangle)} && \text{dep score}
\end{aligned}
$$

Slide from Thadani

joint inference via ILP

objective

$$
\begin{aligned}
C = \arg\max_{\mathbf{x},\mathbf{y},\mathbf{z}} \quad & \sum_i x_i \cdot \mathbf{w}_{tok}^\top \phi(t_i) \\
+ \quad & \sum_{i,j,k} y_{ijk} \cdot \mathbf{w}_{ngr}^\top \phi(\langle t_i, t_j, t_k \rangle) \\
+ \quad & \sum_{i,j} z_{ij} \cdot \mathbf{w}_{dep}^\top \phi(\langle t_i, t_j \rangle)
\end{aligned}
$$

token score

ngram score

dep score

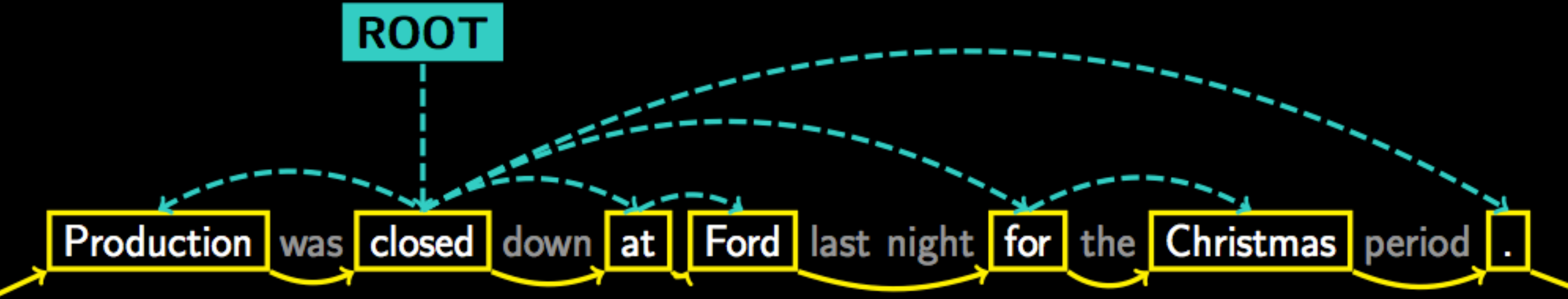
features

- informativeness
- fluency
- fidelity
- pseudo-normalization

Slide from Thadani

Compression

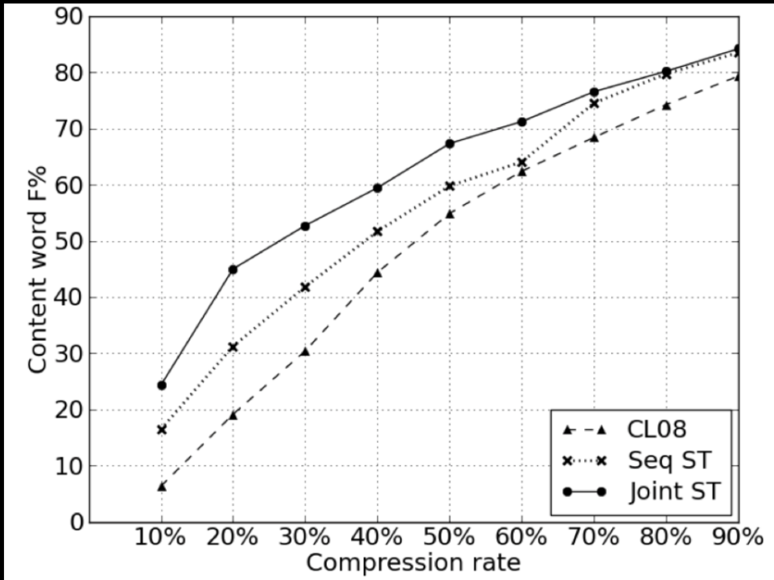
- Input: single sentence
- Output: sentence with **salient** information
- Dataset + baseline from *Clarke & Lapata (2008)*



What Have We Learned?

Compression

+5% n-gram recall for joint inference with dependency relations



Going forward

- Building a corpus for future learning
- Neural net models for *controllable* compression

Multi-Document Summarization Research Focus

- Monitor variety of online information sources
  - News, multilingual
  - Email
- Gather information on events across source and time
  - Same day, multiple sources
  - Across time
- Summarize
  - Highlighting similarities, new information, different perspectives, user specified interests in real-time

Our Approach

- Use a hybrid of statistical and linguistic knowledge
- Statistical analysis of multiple documents
    - Identify important new, contradictory information
- Information fusion
- Generation of summary sentences
    - By re-using phrases
    - Automatic editing/rewriting summary

Newsblaster

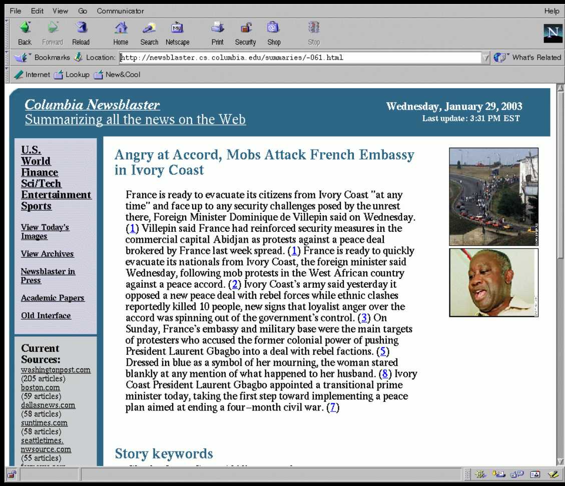
Integrated in online environment for daily news updates

[*http://newsblaster.cs.columbia.edu/*](http://newsblaster.cs.columbia.edu/)

Ani Nenkova

David Elson

Newsblaster



File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://newsblaster.cs.columbia.edu/summaries/-061.html What's Related
Internet Lookup New&Cool

Columbia Newsblaster
Summarizing all the news on the Web

Wednesday, January 29, 2003
Last update: 3:31 PM EST

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images

View Archives

Newsblaster in Press

Academic Papers

Old Interface

Current Sources:
washingtonpost.com
(205 articles)
boston.com
(59 articles)
dallasnews.com
(58 articles)
suntimes.com
(58 articles)
seattletimes.nwsource.com
(55 articles)

Angry at Accord, Mobs Attack French Embassy in Ivory Coast

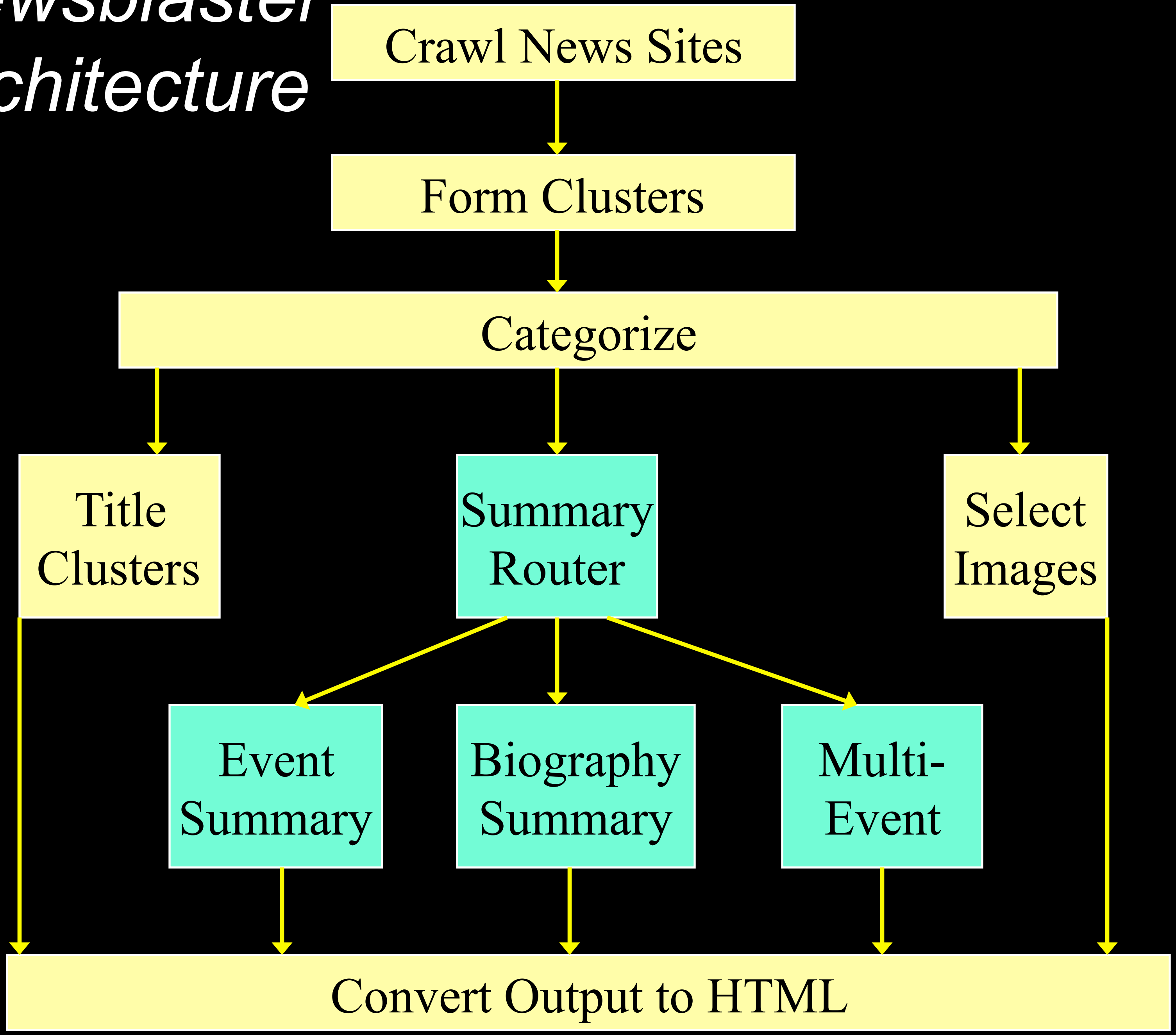
France is ready to evacuate its citizens from Ivory Coast "at any time" and face up to any security challenges posed by the unrest there, Foreign Minister Dominique de Villepin said on Wednesday. (1) Villepin said France had reinforced security measures in the commercial capital Abidjan as protests against a peace deal brokered by France last week spread. (1) France is ready to quickly evacuate its nationals from Ivory Coast, the foreign minister said Wednesday, following mob protests in the West African country against a peace accord. (2) Ivory Coast's army said yesterday it opposed a new peace deal with rebel forces while ethnic clashes reportedly killed 10 people, new signs that loyalist anger over the accord was spinning out of the government's control. (3) On Sunday, France's embassy and military base were the main targets of protesters who accused the former colonial power of pushing President Laurent Gbagbo into a deal with rebel factions. (5) Dressed in blue as a symbol of her mourning, the woman stared blankly at any mention of what happened to her husband. (8) Ivory Coast President Laurent Gbagbo appointed a transitional prime minister today, taking the first step toward implementing a peace plan aimed at ending a four–month civil war. (7)

Story keywords

http://newsblaster.cs.columbia.edu/

- Clustering articles into events
- Categorization by broad topic
- Multi-document summarization
- Generation of summary sentences
  - Fusion
  - Editing of references

Newsblaster Architecture



Crawl News Sites

Form Clusters

Categorize

Title Clusters

Summary Router

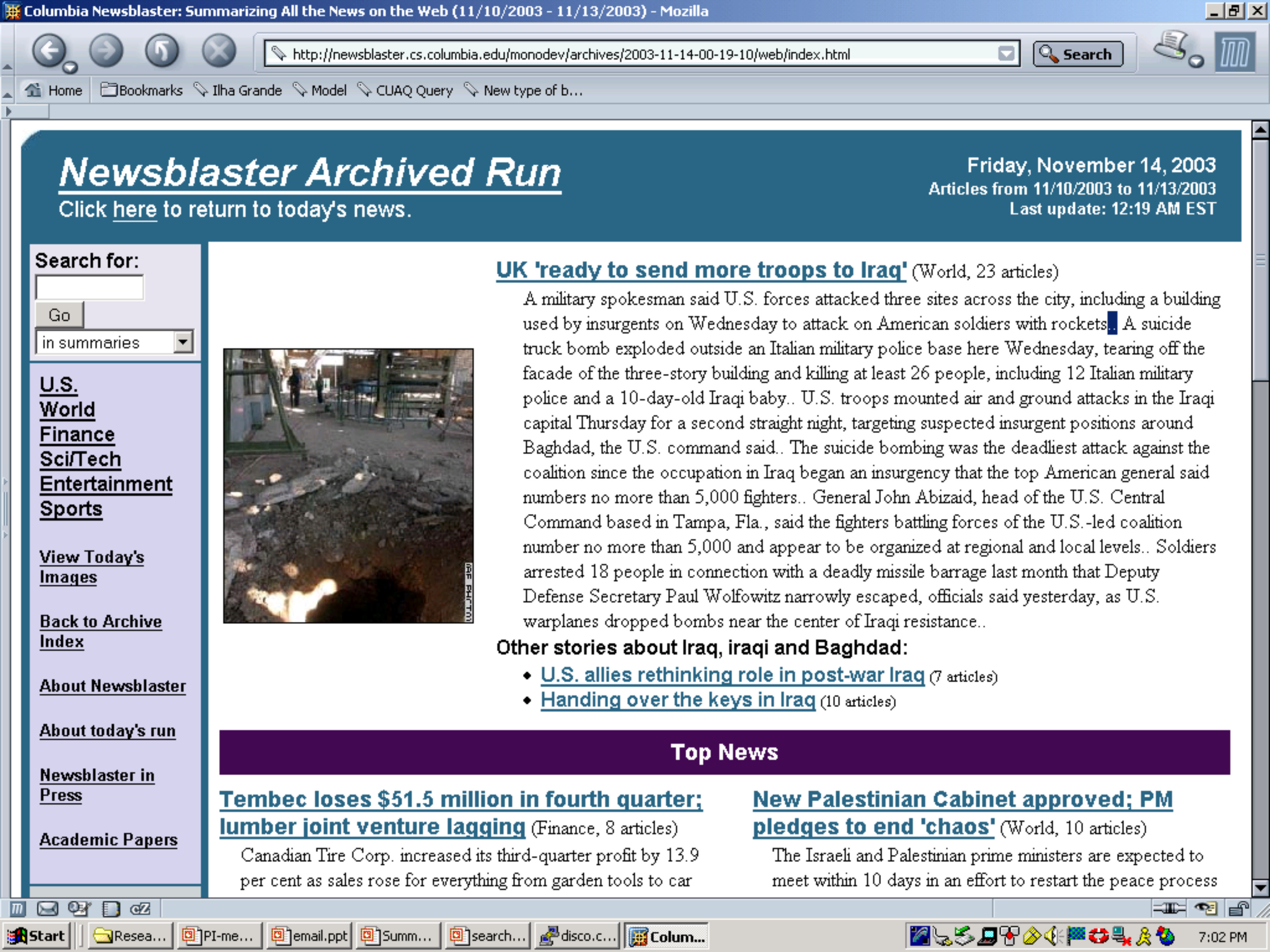
Select Images

Event Summary

Biography Summary

Multi-Event

Convert Output to HTML


Columbia Newsblaster: Summarizing All the News on the Web (11/10/2003 - 11/13/2003) - Mozilla
http://newsblaster.cs.columbia.edu/monodev/archives/2003-11-14-00-19-10/web/index.html
Search
Home
Bookmarks
Ilha Grande
Model
CUAQ Query
New type of b...
Start
Resea...
PI-me...
email.ppt
Summ...
search...
disco.c...
Colum...
7:02 PM

Newsblaster Archived Run

Click here to return to today's news.

Friday, November 14, 2003
Articles from 11/10/2003 to 11/13/2003
Last update: 12:19 AM EST

Search for:

Go

in summaries

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images

Back to Archive Index

About Newsblaster

About today's run

Newsblaster in Press

Academic Papers

UK 'ready to send more troops to Iraq' (World, 23 articles)

A military spokesman said U.S. forces attacked three sites across the city, including a building used by insurgents on Wednesday to attack on American soldiers with rockets.. A suicide truck bomb exploded outside an Italian military police base here Wednesday, tearing off the facade of the three-story building and killing at least 26 people, including 12 Italian military police and a 10-day-old Iraqi baby.. U.S. troops mounted air and ground attacks in the Iraqi capital Thursday for a second straight night, targeting suspected insurgent positions around Baghdad, the U.S. command said.. The suicide bombing was the deadliest attack against the coalition since the occupation in Iraq began an insurgency that the top American general said numbers no more than 5,000 fighters.. General John Abizaid, head of the U.S. Central Command based in Tampa, Fla., said the fighters battling forces of the U.S.-led coalition number no more than 5,000 and appear to be organized at regional and local levels.. Soldiers arrested 18 people in connection with a deadly missile barrage last month that Deputy Defense Secretary Paul Wolfowitz narrowly escaped, officials said yesterday, as U.S. warplanes dropped bombs near the center of Iraqi resistance..

Other stories about Iraq, iraqi and Baghdad:

- U.S. allies rethinking role in post-war Iraq (7 articles)
- Handing over the keys in Iraq (10 articles)

Top News

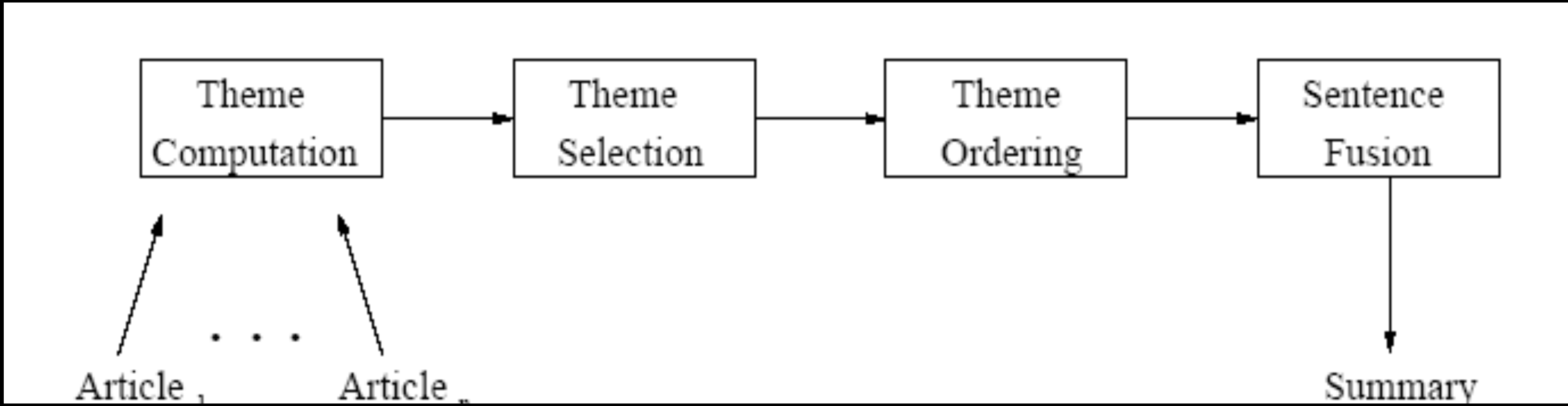
Tembec loses $51.5 million in fourth quarter; lumber joint venture lagging (Finance, 8 articles)

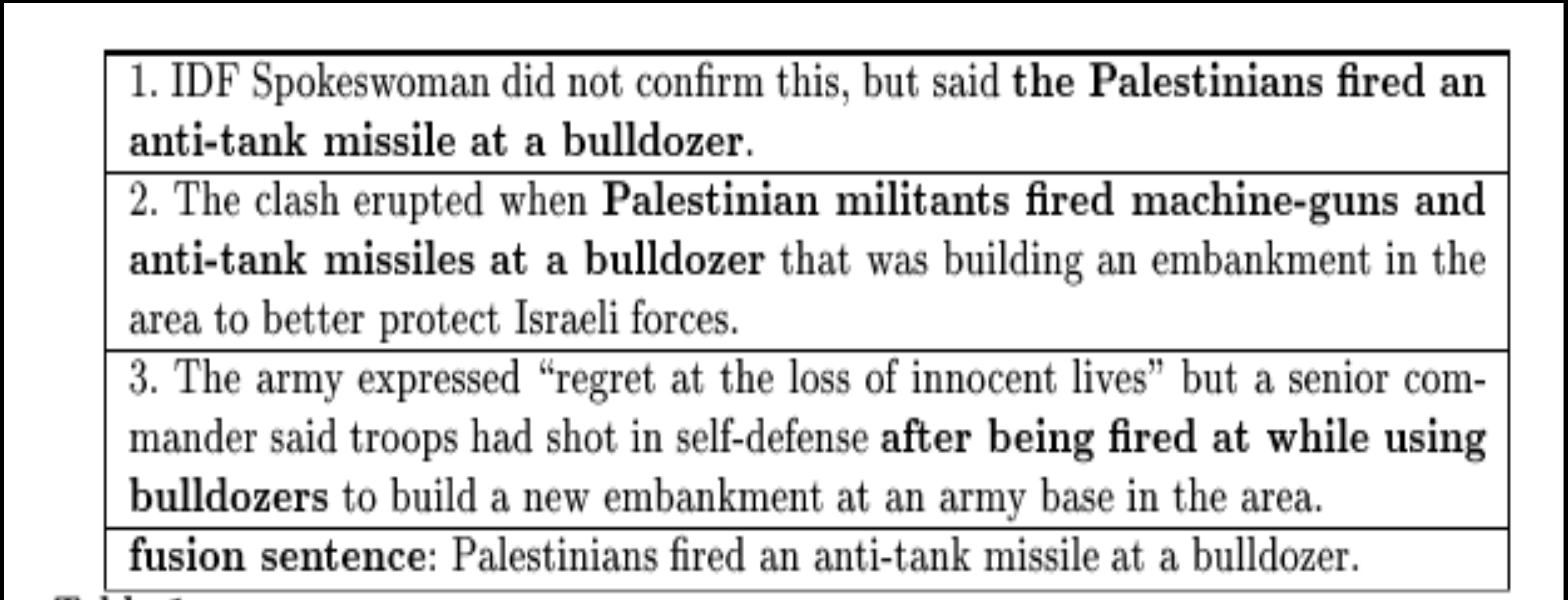
Canadian Tire Corp. increased its third-quarter profit by 13.9 per cent as sales rose for everything from garden tools to car

New Palestinian Cabinet approved; PM pledges to end 'chaos' (World, 10 articles)

The Israeli and Palestinian prime ministers are expected to meet within 10 days in an effort to restart the peace process

Sentence Fusion





1. IDF Spokeswoman did not confirm this, but said **the Palestinians fired an anti-tank missile at a bulldozer**.

2. The clash erupted when **Palestinian militants fired machine-guns and anti-tank missiles at a bulldozer** that was building an embankment in the area to better protect Israeli forces.

3. The army expressed "regret at the loss of innocent lives" but a senior commander said troops had shot in self-defense **after being fired at while using bulldozers** to build a new embankment at an army base in the area.

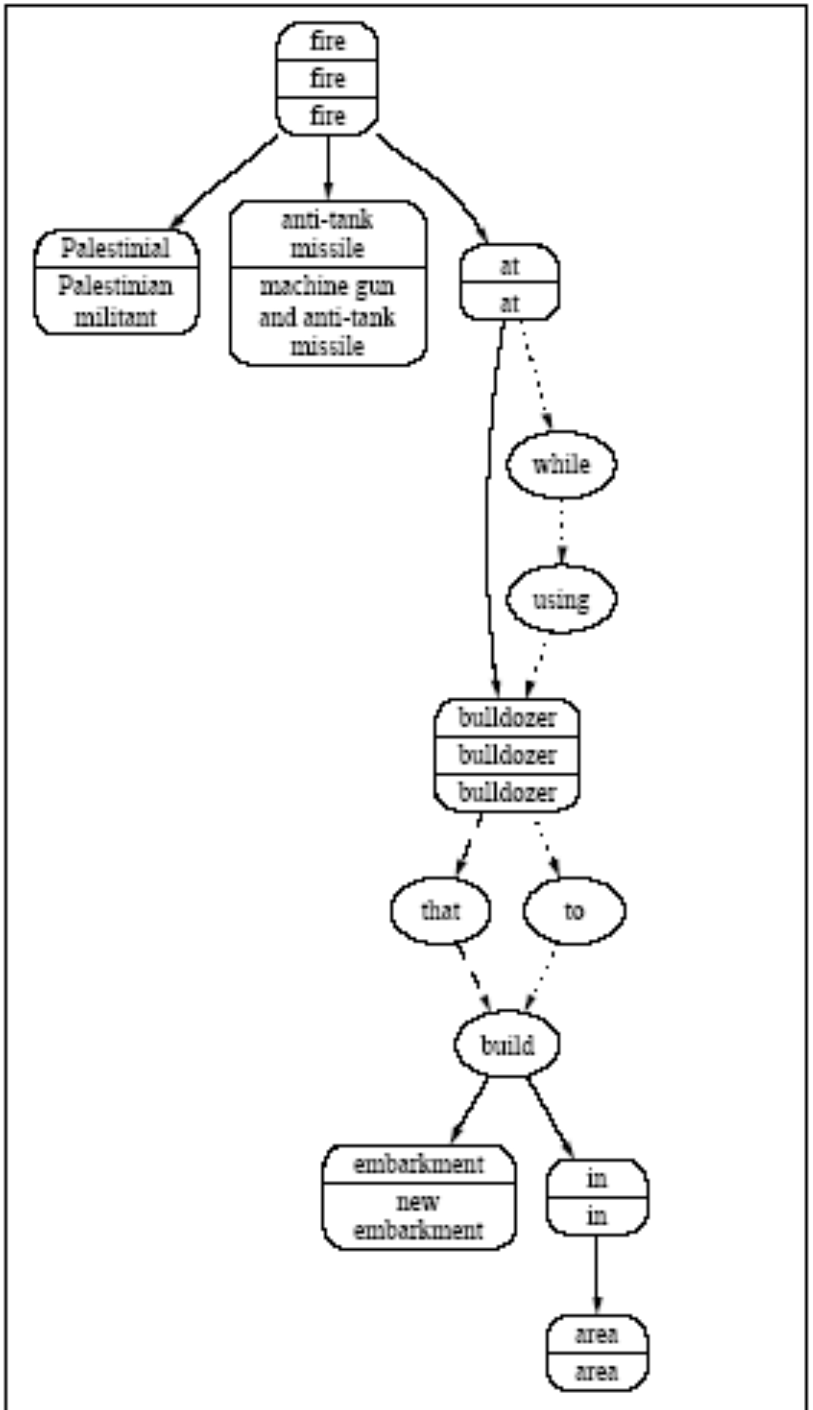
fusion sentence: Palestinians fired an anti-tank missile at a bulldozer.

Theme Computation

- Input: A set of related documents
- Output: Sets of sentences that “mean” the same thing
- Algorithm
  - Compute similarity across sentences using the Cosine Metric
  - Can compare word overlap or phrase overlap
  - IR vector space model could be substituted

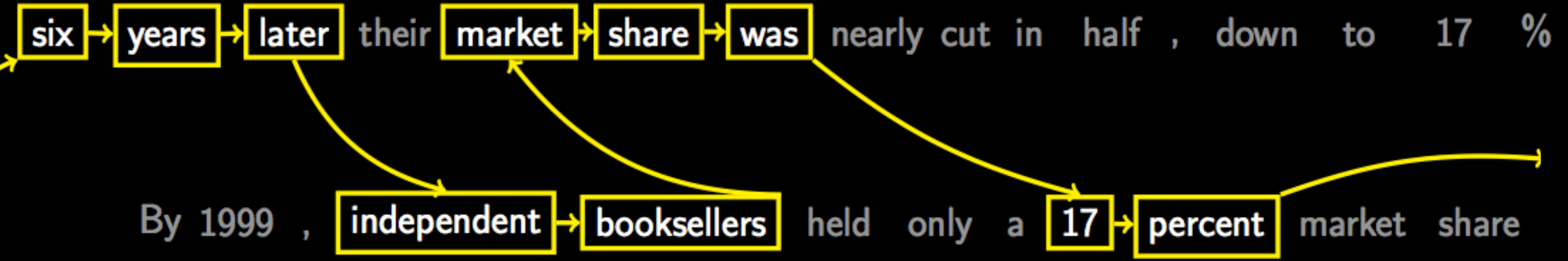
Sentence Fusion Computation

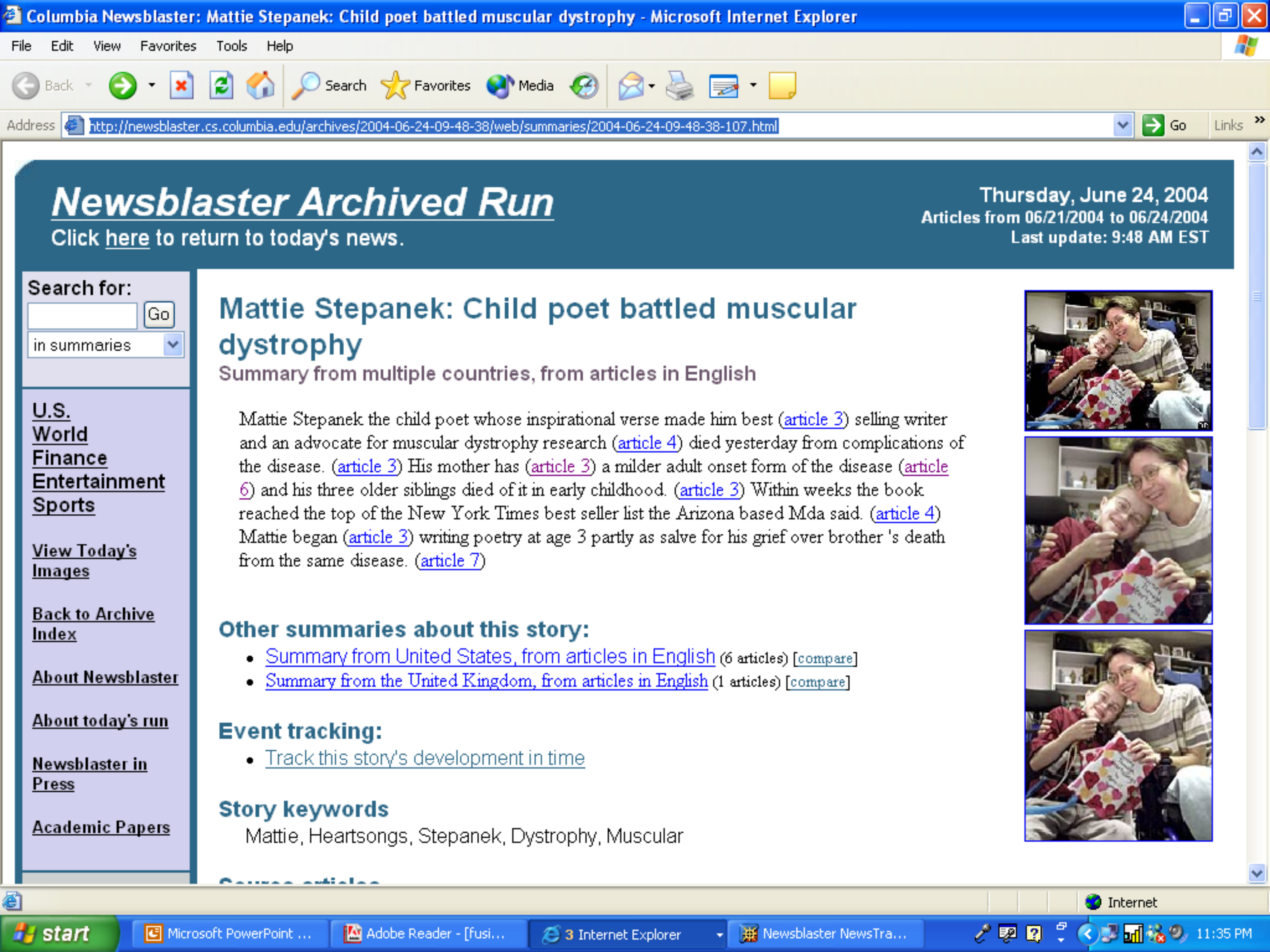
- Common information identification
  - Alignment of constituents in parsed theme sentences: *only some subtrees match*
  - Bottom-up local multi-sequence alignment
  - Similarity depends on
    - Word/paraphrase similarity
    - Tree structure similarity
- Fusion lattice computation
  - Choose a basis sentence
  - Add subtrees from fusion not present in basis
  - Add alternative verbalizations
  - Remove subtrees from basis not present in fusion
- Lattice linearization
  - Generate all possible sentences from the fusion lattice
  - Score sentences using statistical language model


fire
fire
fire
Palestinial
Palestinian militant
anti-tank missile
machine gun and anti-tank missile
at
at
while
using
bulldozer
bulldozer
bulldozer
that
to
build
embarkment
new embarkment
in
in
area
area

Sentence Fusion – Structured Prediction

- Input: multiple sentences
- Output: sentence with ***common*** information
- Dataset created from summarization evaluations
- Fusion-specific features, e.g., repetition


six years later their market share was nearly cut in half , down to 17 %
By 1999 , independent booksellers held only a 17 percent market share



Columbia Newsblaster: Mattie Stepanek: Child poet battled muscular dystrophy - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://newsblaster.cs.columbia.edu/archives/2004-06-24-09-48-38/web/summaries/2004-06-24-09-48-38-107.html
Go
Links

Newsblaster Archived Run

Click here to return to today's news.

Thursday, June 24, 2004
Articles from 06/21/2004 to 06/24/2004
Last update: 9:48 AM EST

Search for:
Go
in summaries
U.S.
World
Finance
Entertainment
Sports
View Today's Images
Back to Archive Index
About Newsblaster
About today's run
Newsblaster in Press
Academic Papers

Mattie Stepanek: Child poet battled muscular dystrophy

Summary from multiple countries, from articles in English

Mattie Stepanek the child poet whose inspirational verse made him best (article 3) selling writer and an advocate for muscular dystrophy research (article 4) died yesterday from complications of the disease. (article 3) His mother has (article 3) a milder adult onset form of the disease (article 6) and his three older siblings died of it in early childhood. (article 3) Within weeks the book reached the top of the New York Times best seller list the Arizona based Mda said. (article 4) Mattie began (article 3) writing poetry at age 3 partly as salve for his grief over brother 's death from the same disease. (article 7)

Other summaries about this story:

- Summary from United States, from articles in English (6 articles) [compare]
- Summary from the United Kingdom, from articles in English (1 articles) [compare]

Event tracking:

- Track this story's development in time

Story keywords

Mattie, Heartsongs, Stepanek, Dystrophy, Muscular

Source articles

What Can Go Wrong?

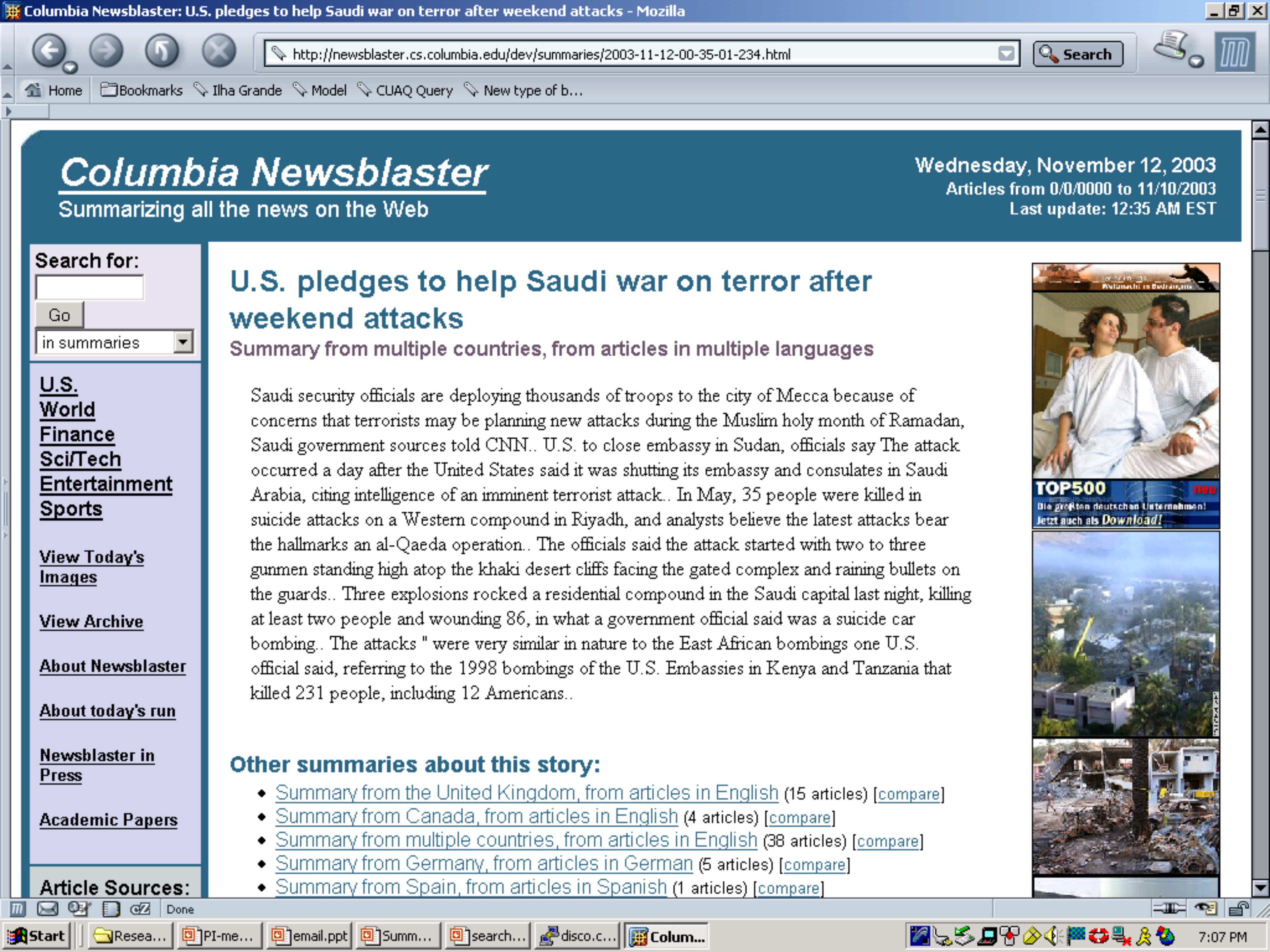
- Family names in the news: who’s who?

What Can Go Wrong?

- Family names in the news: who's who?
- On the death of the Queen Mother in England, Newsblaster had Queen Elizabeth attending her own funeral.

Different Perspectives

- Hierarchical clustering
    - Each event cluster is divided into clusters by country
- Different perspectives can be viewed side by side
- Experimenting with update summarizer to identify key differences between sets of stories


Columbia Newsblaster: U.S. pledges to help Saudi war on terror after weekend attacks - Mozilla
http://newsblaster.cs.columbia.edu/dev/summaries/2003-11-12-00-35-01-234.html
Search
Home
Bookmarks
Ilha Grande
Model
CUAQ Query
New type of b...
TOP500
Die größten deutschen Unternehmen!
Jetzt auch als Download!
Done
Start
Resea...
PI-me...
email.ppt
Summ...
search...
disco.c...
Colum...
7:07 PM

Columbia Newsblaster
Summarizing all the news on the Web

Wednesday, November 12, 2003
Articles from 0/0/0000 to 11/10/2003
Last update: 12:35 AM EST

Search for:

Go

in summaries

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images

View Archive

About Newsblaster

About today's run

Newsblaster in Press

Academic Papers

Article Sources:

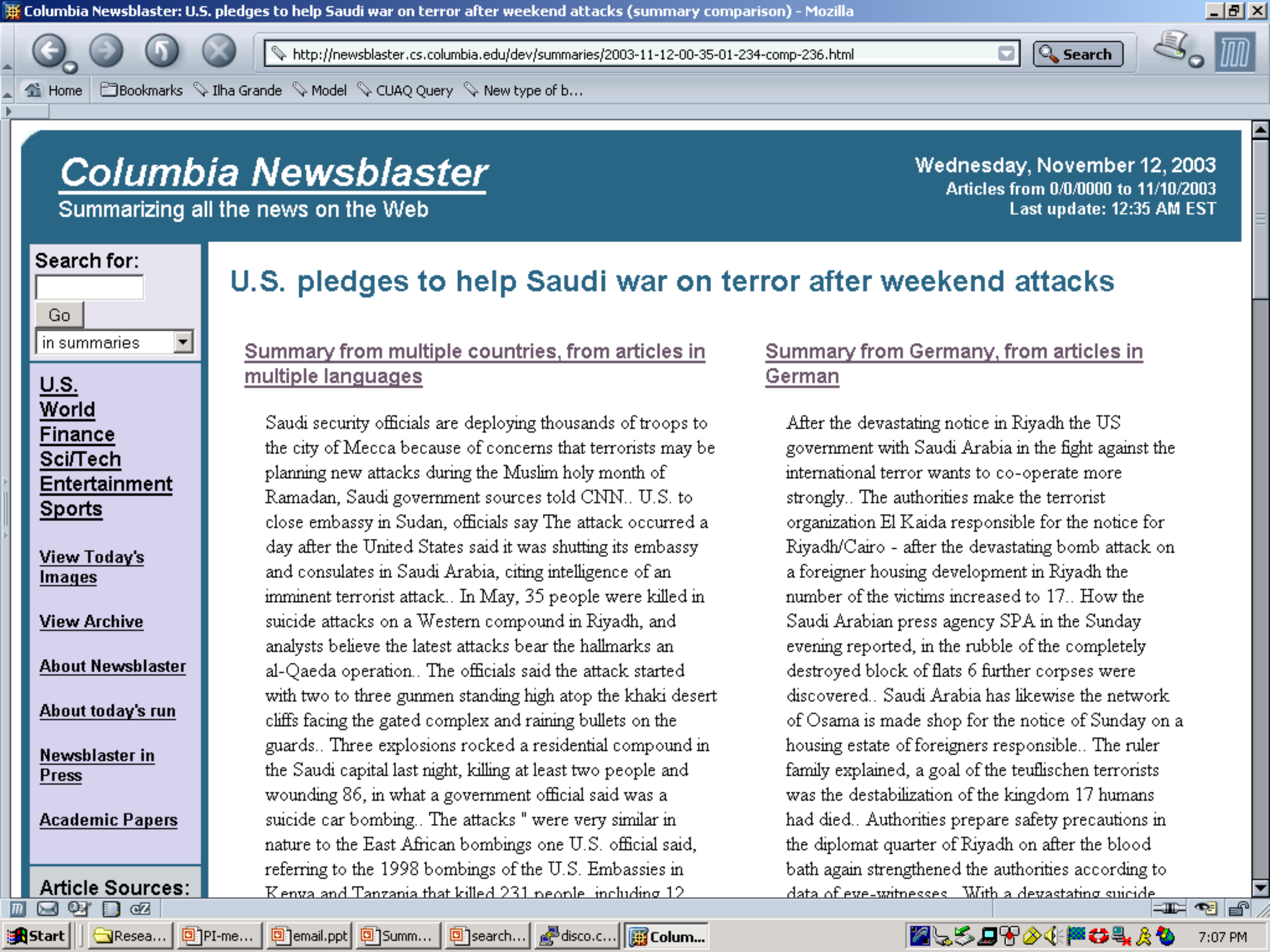
U.S. pledges to help Saudi war on terror after weekend attacks

Summary from multiple countries, from articles in multiple languages

Saudi security officials are deploying thousands of troops to the city of Mecca because of concerns that terrorists may be planning new attacks during the Muslim holy month of Ramadan, Saudi government sources told CNN.. U.S. to close embassy in Sudan, officials say The attack occurred a day after the United States said it was shutting its embassy and consulates in Saudi Arabia, citing intelligence of an imminent terrorist attack.. In May, 35 people were killed in suicide attacks on a Western compound in Riyadh, and analysts believe the latest attacks bear the hallmarks an al-Qaeda operation.. The officials said the attack started with two to three gunmen standing high atop the khaki desert cliffs facing the gated complex and raining bullets on the guards.. Three explosions rocked a residential compound in the Saudi capital last night, killing at least two people and wounding 86, in what a government official said was a suicide car bombing.. The attacks " were very similar in nature to the East African bombings one U.S. official said, referring to the 1998 bombings of the U.S. Embassies in Kenya and Tanzania that killed 231 people, including 12 Americans..

Other summaries about this story:

- Summary from the United Kingdom, from articles in English (15 articles) [compare]
- Summary from Canada, from articles in English (4 articles) [compare]
- Summary from multiple countries, from articles in English (38 articles) [compare]
- Summary from Germany, from articles in German (5 articles) [compare]
- Summary from Spain, from articles in Spanish (1 articles) [compare]


Columbia Newsblaster: U.S. pledges to help Saudi war on terror after weekend attacks (summary comparison) - Mozilla
http://newsblaster.cs.columbia.edu/dev/summaries/2003-11-12-00-35-01-234-comp-236.html
Search
Home
Bookmarks
Ilha Grande
Model
CUAQ Query
New type of b...

Columbia Newsblaster

Summarizing all the news on the Web

Wednesday, November 12, 2003
Articles from 0/0/0000 to 11/10/2003
Last update: 12:35 AM EST

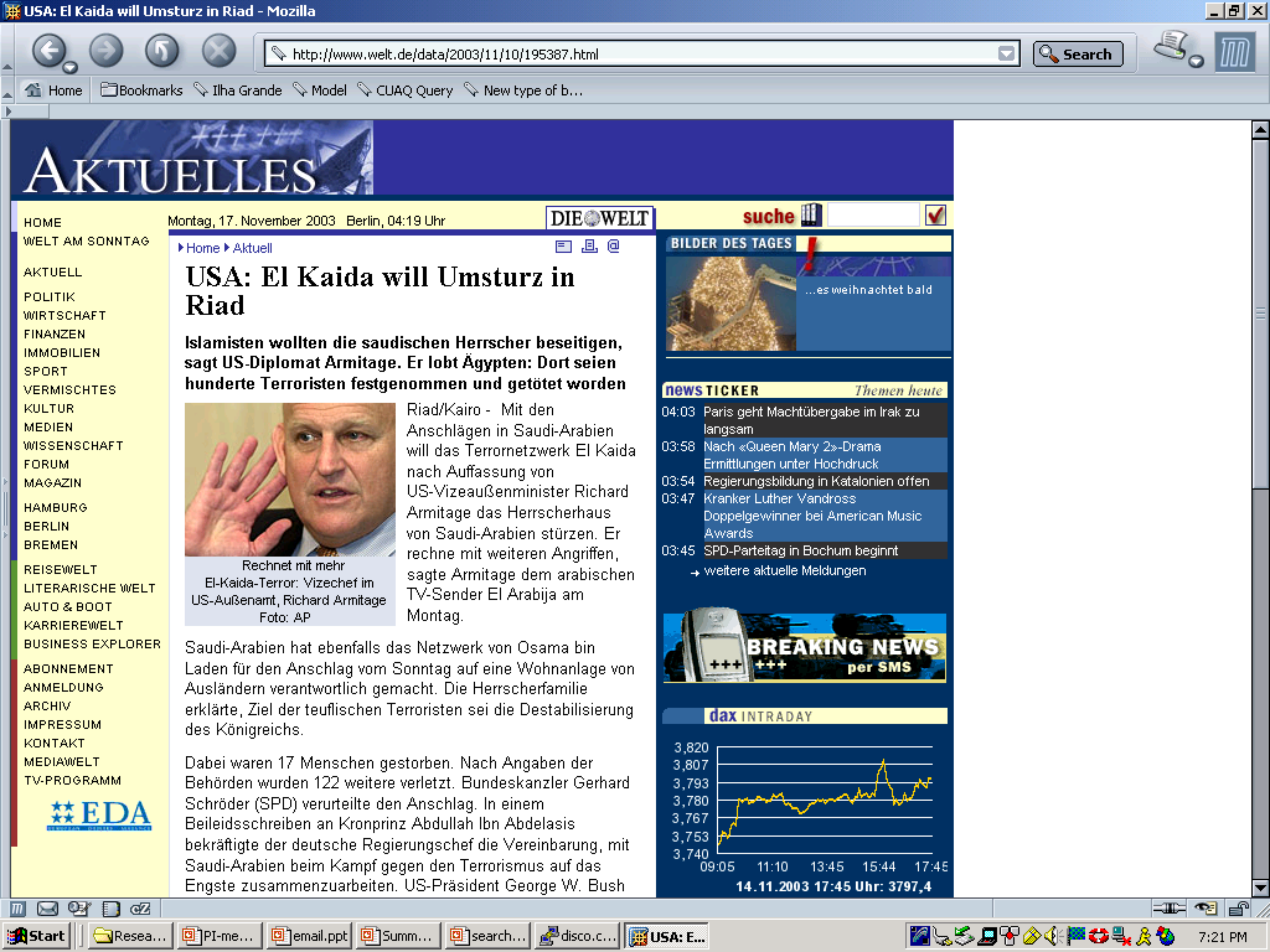
U.S. pledges to help Saudi war on terror after weekend attacks

Summary from multiple countries, from articles in multiple languages

Saudi security officials are deploying thousands of troops to the city of Mecca because of concerns that terrorists may be planning new attacks during the Muslim holy month of Ramadan, Saudi government sources told CNN.. U.S. to close embassy in Sudan, officials say The attack occurred a day after the United States said it was shutting its embassy and consulates in Saudi Arabia, citing intelligence of an imminent terrorist attack.. In May, 35 people were killed in suicide attacks on a Western compound in Riyadh, and analysts believe the latest attacks bear the hallmarks an al-Qaeda operation.. The officials said the attack started with two to three gunmen standing high atop the khaki desert cliffs facing the gated complex and raining bullets on the guards.. Three explosions rocked a residential compound in the Saudi capital last night, killing at least two people and wounding 86, in what a government official said was a suicide car bombing.. The attacks " were very similar in nature to the East African bombings one U.S. official said, referring to the 1998 bombings of the U.S. Embassies in Kenya and Tanzania that killed 231 people, including 12

Summary from Germany, from articles in German

After the devastating notice in Riyadh the US government with Saudi Arabia in the fight against the international terror wants to co-operate more strongly.. The authorities make the terrorist organization El Kaida responsible for the notice for Riyadh/Cairo - after the devastating bomb attack on a foreigner housing development in Riyadh the number of the victims increased to 17.. How the Saudi Arabian press agency SPA in the Sunday evening reported, in the rubble of the completely destroyed block of flats 6 further corpses were discovered.. Saudi Arabia has likewise the network of Osama is made shop for the notice of Sunday on a housing estate of foreigners responsible.. The ruler family explained, a goal of the teuflischen terrorists was the destabilization of the kingdom 17 humans had died.. Authorities prepare safety precautions in the diplomat quarter of Riyadh on after the blood bath again strengthened the authorities according to data of eye-witnesses.. With a devastating suicide



Montag, 17. November 2003 Berlin, 04:19 Uhr

Home ▸ Aktuell

USA: El Kaida will Umsturz in Riad

Islamisten wollten die saudischen Herrscher beseitigen, sagt US-Diplomat Armitage. Er lobt Ägypten: Dort seien hunderte Terroristen festgenommen und getötet worden

Rechnet mit mehr El-Kaida-Terror: Vizechef im US-Außenamt, Richard Armitage
Foto: AP

Riad/Kairo - Mit den Anschlägen in Saudi-Arabien will das Terrornetzwerk El Kaida nach Auffassung von US-Vizeaußenminister Richard Armitage das Herrscherhaus von Saudi-Arabien stürzen. Er rechne mit weiteren Angriffen, sagte Armitage dem arabischen TV-Sender El Arabija am Montag.

Saudi-Arabien hat ebenfalls das Netzwerk von Osama bin Laden für den Anschlag vom Sonntag auf eine Wohnanlage von Ausländern verantwortlich gemacht. Die Herrscherfamilie erklärte, Ziel der teuflischen Terroristen sei die Destabilisierung des Königreichs.

Dabei waren 17 Menschen gestorben. Nach Angaben der Behörden wurden 122 weitere verletzt. Bundeskanzler Gerhard Schröder (SPD) verurteilte den Anschlag. In einem Beileidsschreiben an Kronprinz Abdullah Ibn Abdelasis bekräftigte der deutsche Regierungschef die Vereinbarung, mit Saudi-Arabien beim Kampf gegen den Terrorismus auf das Engste zusammenzuarbeiten. US-Präsident George W. Bush

04:03 Paris geht Machtübergabe im Irak zu langsam
03:58 Nach «Queen Mary 2»-Drama Ermittlungen unter Hochdruck
03:54 Regierungsbildung in Katalonien offen
03:47 Kranker Luther Vandross Doppelgewinner bei American Music Awards
03:45 SPD-Parteitag in Bochum beginnt
→ weitere aktuelle Meldungen

14.11.2003 17:45 Uhr: 3797,4

Multilingual Summarization

- Given a set of documents on the same event
- Some documents are in English
- Some documents are translated from other languages

Issues for Multilingual Summarization

- Problem: Translated text is errorful
- Exploit information available during summarization
    - Similar documents in cluster
- Replace translated sentences with similar English
- Edit translated text
    - Replace named entities with extractions from similar English

Multilingual Redundancy

BAGDAD. - A total of 21 prisoners has been died and a hundred more hurt by firings from mortar in the jail of Abu Gharib (to 20 kilometers to the west of Bagdad), according to has informed general into the U.S.A. Marco Kimmitt.

Spanish

Bagdad in the Iraqi capital Aufstaendi attacked Bagdad on Tuesday a prison with mortars and ***killed after USA gifts 22 prisoners***. Further 92 passengers of the Abu Ghraib prison were hurt, communicated a spokeswoman of the American armed forces.

German

The Iraqi being stationed US military shot on the 20th, the same day to the allied forces detention facility which is in アブグレアブグレイブhdad west approximately 20 kilometers, mortar 12 shot and you were packed, ***22 Iraqi human prisoners died***, it announced that nearly 100 people were injured.

Japanese

BAGHDAD, Iraq – Insurgents fired 12 mortars into Baghdad's Abu Ghraib prison Tuesday, ***killing 22 detainees*** and injuring 92, U.S. military officials said.

English

Multilingual Redundancy

BAGDAD. - A total of 21 prisoners has been died and a hundred more hurt by firings from ***mortar in the jail of Abu Gharib*** (to 20 kilometers to the west of Bagdad), according to has informed general into the U.S.A. Marco Kimmitt.

Spanish

Bagdad   in the Iraqi capital Aufstaendi attacked Bagdad on Tuesday ***a prison with mortars*** and ***killed after USA gifts 22 prisoners***. Further 92 passengers of the Abu Ghraib prison were hurt, communicated a spokeswoman of the American armed forces.

German

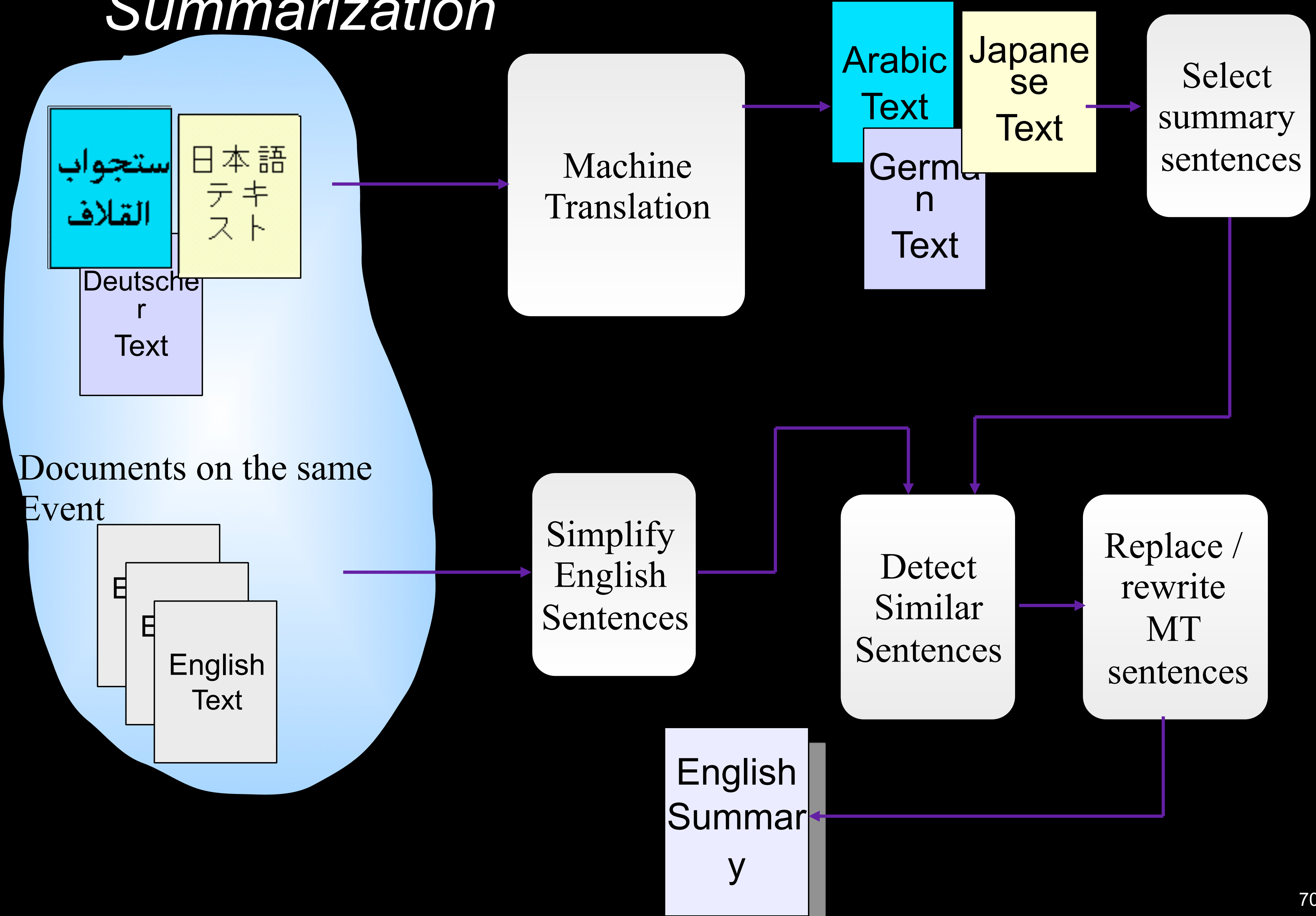
The Iraqi being stationed US military shot on the 20th, the same day to the allied forces detention facility which is in アブグレアブグレイブhdad west approximately 20 kilometers, ***mortar 12 shot*** and you were packed, ***22 Iraqi human prisoners died***, it announced that nearly 100 people were injured.

Japanese

BAGHDAD, Iraq – ***Insurgents fired 12 mortars into Baghdad's Abu Ghraib prison*** Tuesday, ***killing 22 detainees*** and injuring 92, U.S. military officials said.

English

Multilingual Similarity-based Summarization



ستجواب القلاف

日本語 テキスト

Deutscher Text

Documents on the same Event

English Text

Machine Translation

Arabic Text

Japanese Text

German Text

Select summary sentences

Simplify English Sentences

Detect Similar Sentences

Replace / rewrite MT sentences

English Summary

Sentence 1

Iraqi President Saddam Hussein that the government of Iraq over 24 years in a "black" near the port of the northern Iraq after nearly eight months of pursuit was considered the largest in history .

Similarity 0.27: Ousted Iraqi President Saddam Hussein is in custody following his dramatic capture by US forces in Iraq.

Similarity 0.07: Saddam Hussein, the former president of Iraq, has been captured and *is being held by US forces in the country.*

Similarity 0.04: *Coalition authorities have said that the former Iraqi president could be tried at a war crimes tribunal, with Iraqi judges presiding and international legal experts acting as advisers.*

Rewrite proper and common nouns to remove MT errors

(Siddharthan and McKeown 05)

- Use redundancy in input to summarization and multiple translations to build attribute value matrices (AVMs)
  - Record country, role, description for all people
  - Record name variants
- Use generation grammar with semantic categories (role, organization, location) to re-order phrases for fluent output

the representative of Iraq in the United Nations Nizar Hamdoon

+

representative of Iraq of the United Nations Nizar HAMDOON

↓

$$\left[\begin{array}{ll}
\texttt{name} & \text{Nizar Hamdoon} \\
\texttt{role} & \text{representative} \\
\texttt{country} & \text{Iraq } (arg1) \\
\texttt{organization} & \text{United Nations } (arg2)
\end{array}\right]$$

↓

Iraqi United Nations representative Nizar Hamdoon

some likely problems with this approach to NPs?

Current work

- Columbia front page:
- http://www.columbia.edu

Evaluation

- DUC (Document Understanding Conference): run by NIST yearly
- Manual creation of topics (sets of documents)
- 2-7 human written summaries per topic
- How well does a system generated summary cover the information in a human summary?
- Metrics
  - Rouge
  - Pyramid

Rouge

- ROUGE
  - Publicly available at: http://www.isi.edu/~cyl/ROUGE
  - Version 1.2.1 includes:
    - ROUGE-N - n-gram-based co-occurrence statistics
    - ROUGE-L - longest common subsequence-based (LCS) co-occurrence statistics
    - ROUGE-W - LCS-based co-occurrence statistics favoring consecutive LCSes
- Measures recall
    - Rouge-1: How many unigrams in the human summary did the system summary find?
    - Rouge-2: How many bigrams?

Pros and Cons

- Pros
  - Automatic metric: Can be used for tuning
  - With enough examples or enough human models, differences are significant
- Cons
  - In practice, there often aren’t enough examples
  - Measures word overlap so re-wording a problem

Pyramids

- Uses multiple human summaries
    - Previous data indicated 5 needed for score stability
- Information is ranked by its importance
- Allows for multiple good summaries
- A pyramid is created from the human summaries
    - Elements of the pyramid are content units
    - System summaries are scored by comparison with the pyramid

Summarization Content Units

- Near-paraphrases from different human summaries
- Clause or less
- Avoids explicit semantic representation
- Emerges from analysis of human summaries

SCU: A cable car caught fire (Weight = 4)

A. The cause of the fire was unknown.

B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.

C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.

D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

SCU: The cause of the fire is unknown (Weight = 1)

A. The cause of the fire was unknown.

B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.

C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.

D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

SCU: The accident happened in the Austrian Alps (Weight = 3)

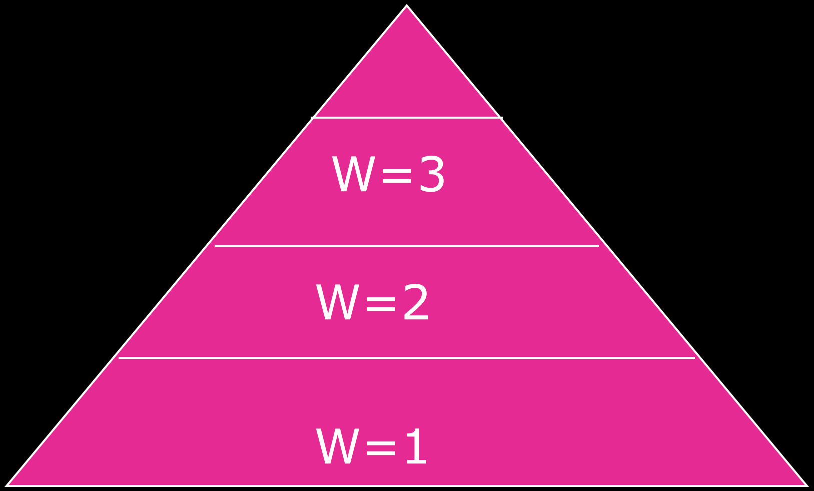
A. The cause of the fire was unknown.

B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.

C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.

D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

Idealized representation



- Tiers of differentially weighted SCUs
- Top: few SCUs, high weight
- Bottom: many SCUs, low weight

Pyramid Score

SCORE = D/MAX

D: Sum of the weights of the SCUs in a summary

MAX: Sum of the weights of the SCUs in a ideally informative summary

Measures the proportion of good information in the summary: precision

User Study: Objectives

- Does multi-document summarization help?
    - Do summaries help the user find information needed to perform a report writing task?
    - Do users use information from summaries in gathering their facts?
    - Do summaries increase user satisfaction with the online news system?
    - Do users create better quality reports with summaries?
    - How do full multi-document summaries compare with minimal 1-sentence summaries such as Google News?

User Study: Design

- Four parallel news systems
  - *Source documents only;* no summaries
  - *Minimal single sentence summaries* (Google News)
  - *Newsblaster summaries*
  - *Human summaries*
- All groups write reports given four scenarios
  - A task similar to analysts
  - Can only use Newsblaster for research
  - Time-restricted

User Study: Execution

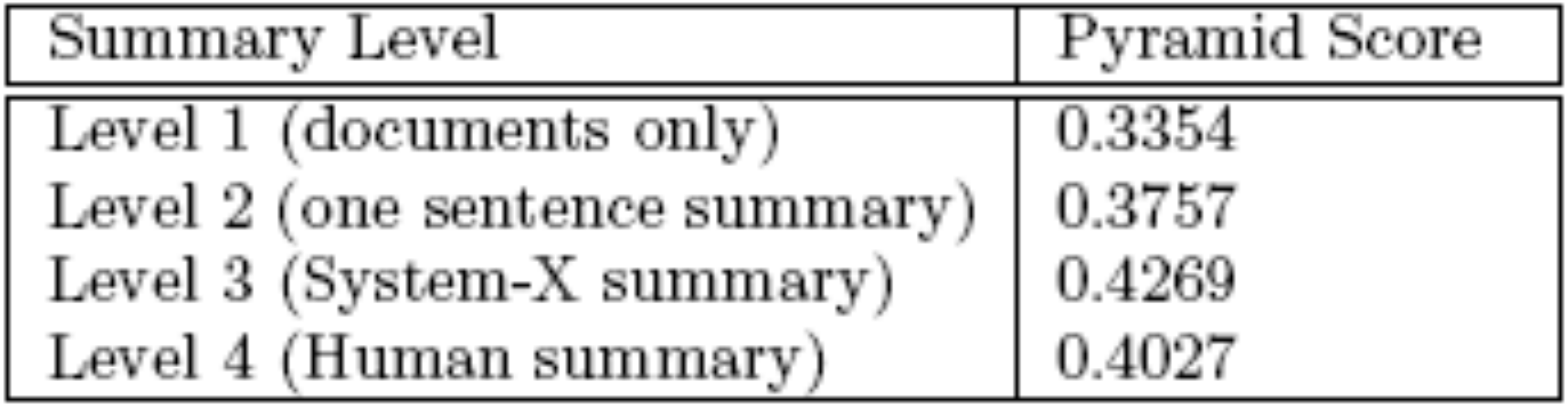
- 4 scenarios
  - 4 event clusters each
  - 2 directly relevant, 2 peripherally relevant
  - Average 10 documents/cluster
- 45 participants
  - Balance between liberal arts, engineering
  - 138 reports
- Exit survey
  - Multiple-choice and open-ended questions
- Usage tracking
  - Each click logged, on or off-site

“Geneva” Prompt

- The conflict between Israel and the Palestinians has been difficult for government negotiators to settle. Most recently, implementation of the “road map for peace”, a diplomatic effort sponsored by ……
  - Who participated in the negotiations that produced the Geneva Accord?
  - Apart from direct participants, who supported the Geneva Accord preparations and how?
  - What has the response been to the Geneva Accord by the Palestinians?

Measuring Effectiveness

- Score report content and compare across summary conditions
- Compare user satisfaction per summary condition
- Comparing where subjects took report content from



Summary Level	Pyramid Score
Level 1 (documents only)	0.3354
Level 2 (one sentence summary)	0.3757
Level 3 (System-X summary)	0.4269
Level 4 (Human summary)	0.4027

Table 2: Mean Pyramid Scores on Reports, Scenario 1 (Geneva Accords) excluded.

User Satisfaction

- More effective than a web search with Newsblaster
    - Not true with documents only or single-sentence summaries
- Easier to complete the task with summaries than with documents only
- Enough time with summaries than documents only
- Summaries helped most
    - 5% single sentence summaries
    - 24% Newsblaster summaries
    - 43% human summaries

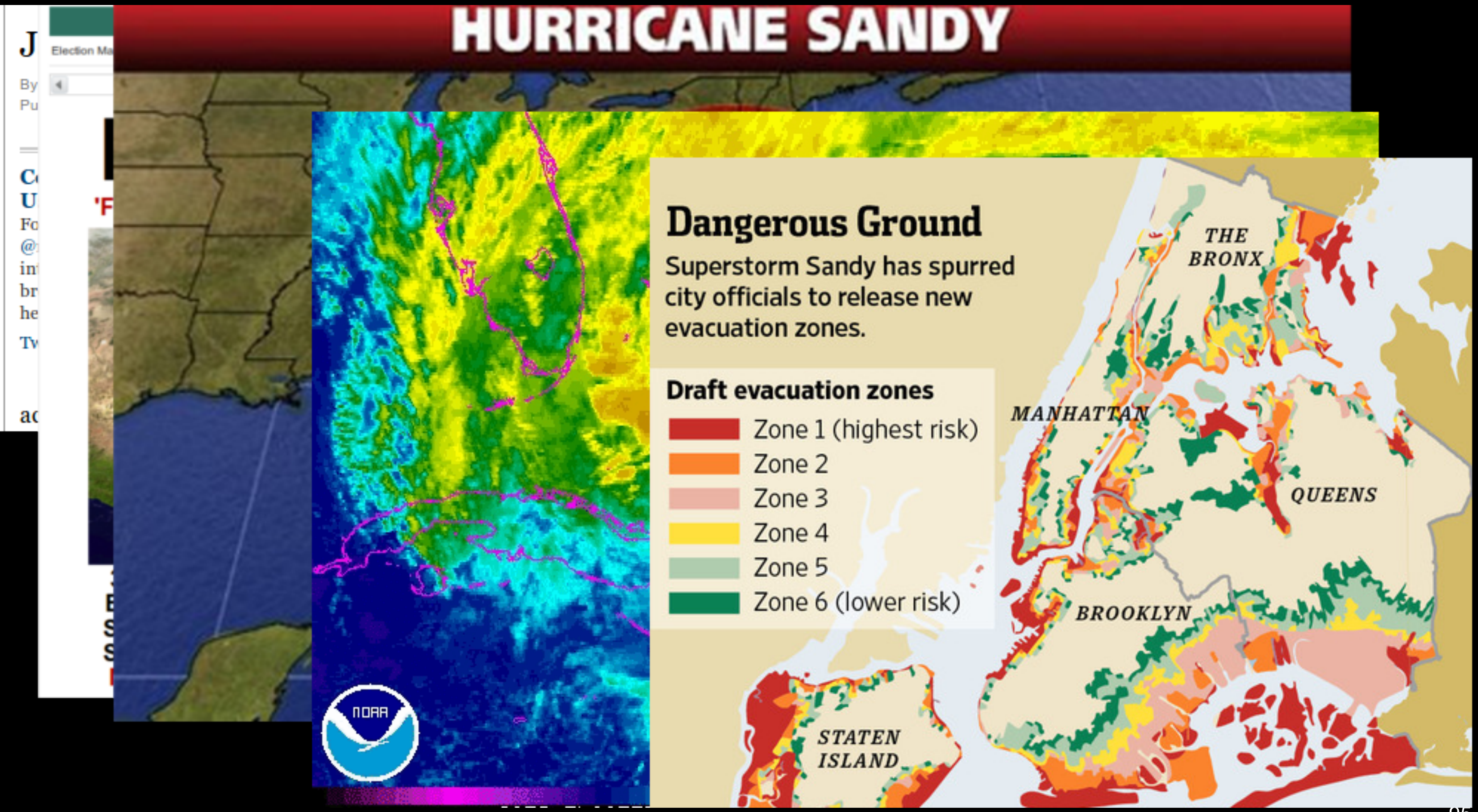
User Study: Conclusions

- Summaries measurably improve a news browswer’s effectiveness for research
- Users are more satisfied with Newsblaster summaries are better than single-sentence summaries like those of Google News
- Users want search
  - Not included in evaluation

Questions (from Sparck Jones)

- Should we take the reader into account and how?
- Need more power than text extraction and more flexibility than fact extraction (p. 4)
- “Similarly, the notion of a basic summary, i.e., one reflective of the source, makes hidden fact assumptions, for example that the subject knowledge of the output’s readers will be on a par with that of the readers for whom the source was intended. (p. 5)”
- Is the state of the art sufficiently mature to allow summarization from intermediate representations and still allow robust processing of domain independent material?
- Evaluation: gold standard vs. user study? Difficulty of evaluation?

Problem: Identifying needs during disaster



Monitor events over time

- Input: streaming data
- News, web pages
- At every hour, what’s new

Track events and SubEvents

Hurricane Sandy

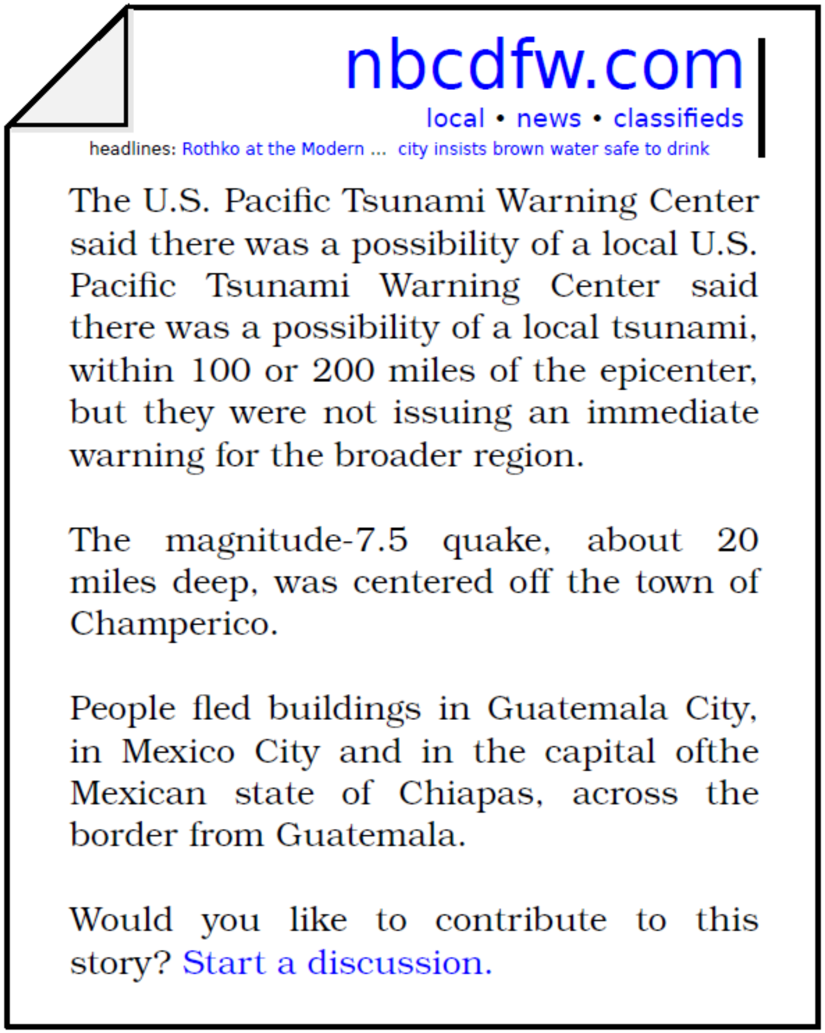
Manhattan Blackout

Breezy Point fire

Public Transit Outage

Data from NIST: 2011 – 2013
Web Crawl, 11 categories

:15



nbcdfw.com

local • news • classifieds

headlines: Rothko at the Modern ... city insists brown water safe to drink

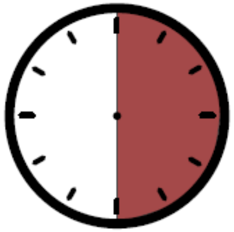
The U.S. Pacific Tsunami Warning Center said there was a possibility of a local U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.

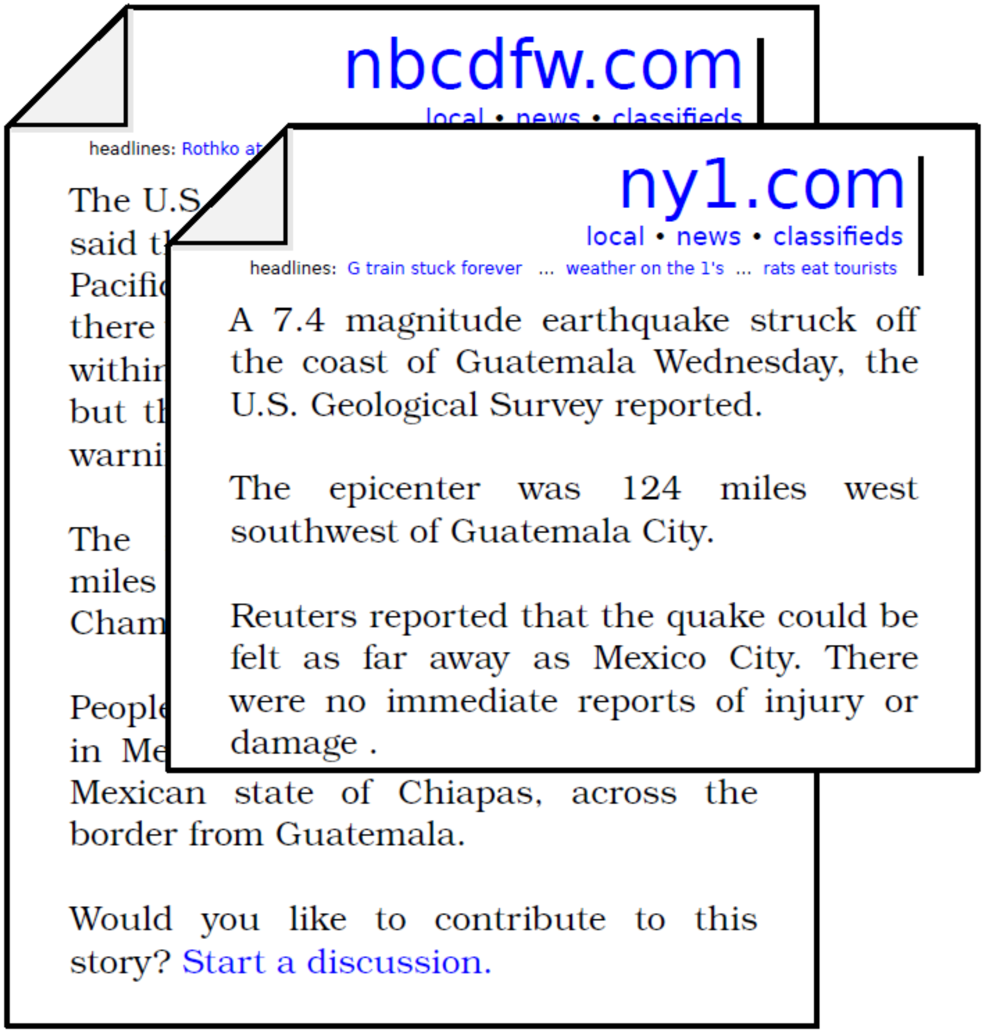
The magnitude-7.5 quake, about 20 miles deep, was centered off the town of Champerico.

People fled buildings in Guatemala City, in Mexico City and in the capital ofthe Mexican state of Chiapas, across the border from Guatemala.

Would you like to contribute to this story? Start a discussion.

:30





nbcdfw.com

local • news • classifieds

headlines: Rothko at

The U.S
said t
Pacific
there
within
but th
warni

The
miles
Cham

People
in Me
Mexican state of Chiapas, across the border from Guatemala.

Would you like to contribute to this story? Start a discussion.

ny1.com

local • news • classifieds

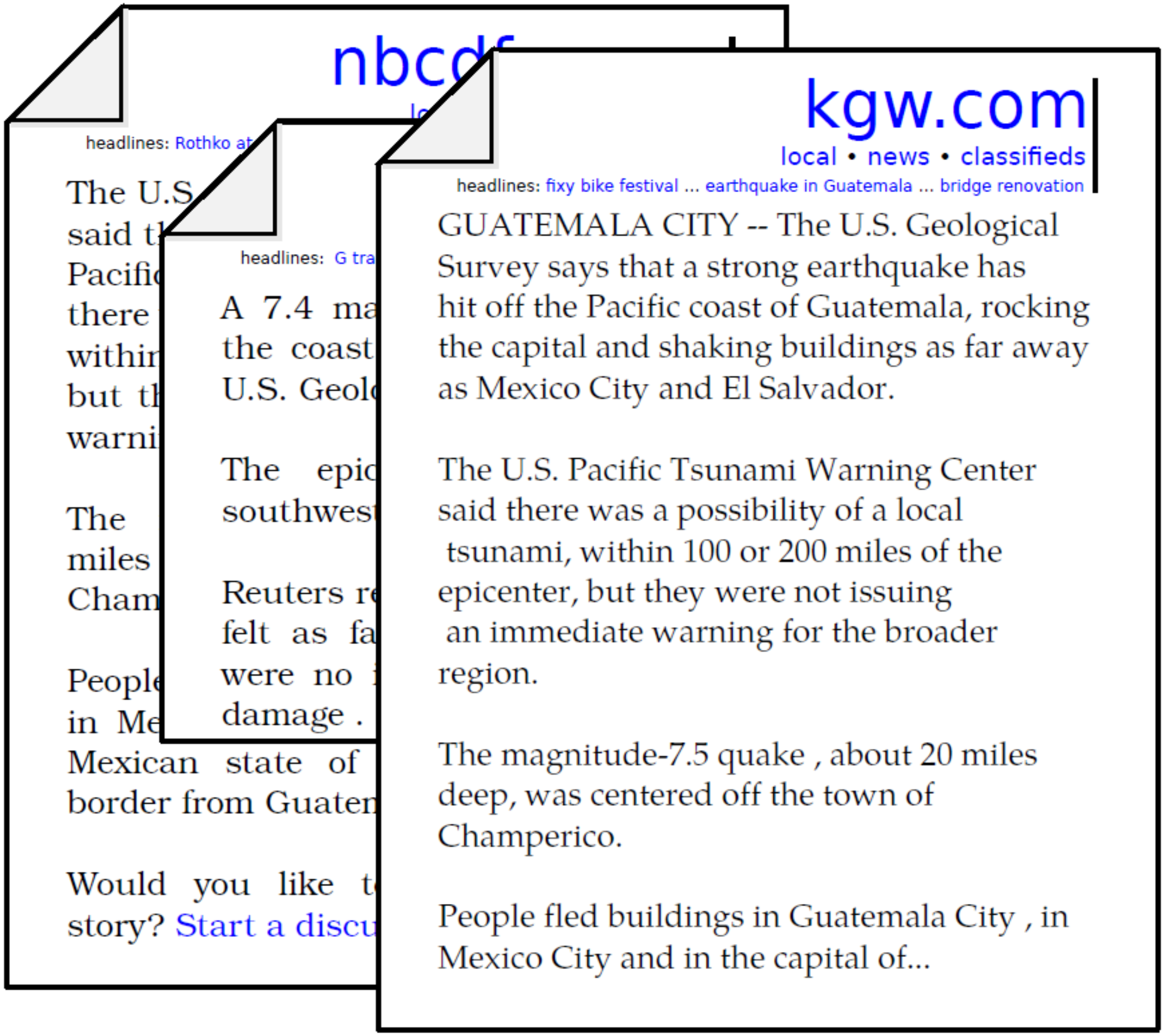
headlines: G train stuck forever ... weather on the 1's ... rats eat tourists

A 7.4 magnitude earthquake struck off the coast of Guatemala Wednesday, the U.S. Geological Survey reported.

The epicenter was 124 miles west southwest of Guatemala City.

Reuters reported that the quake could be felt as far away as Mexico City. There were no immediate reports of injury or damage .

:45



nbcd

headlines: Rothko at

The U.S
said t
Pacifi
there
withi
but t
warni

The
miles
Cham

People
in Me
Mexican state of
border from Guate

Would you like t
story? Start a discu

headlines: G tra

A 7.4 ma
the coast
U.S. Geol

The epic
southwest

Reuters r
felt as fa
were no
damage .

kgw.com

local • news • classifieds

headlines: fixy bike festival ... earthquake in Guatemala ... bridge renovation

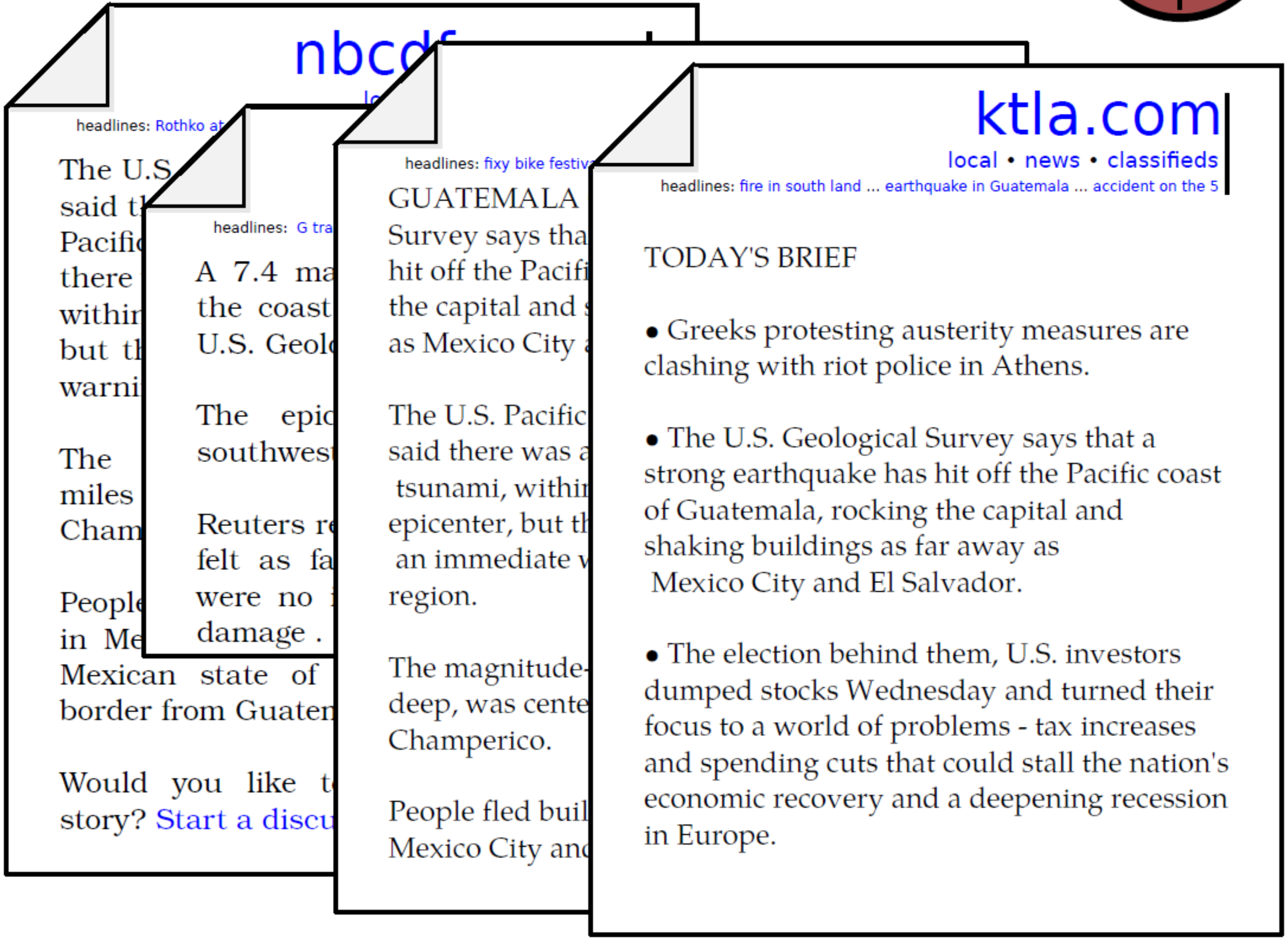
GUATEMALA CITY -- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.

The U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.

The magnitude-7.5 quake , about 20 miles deep, was centered off the town of Champerico.

People fled buildings in Guatemala City , in Mexico City and in the capital of...

1:00



nbcdf

lo

headlines: Rothko at

The U.S
said t
Pacific
there
within
but th
warni

The
miles
Cham

People
in Me
Mexican state of
border from Guaten

Would you like t
story? Start a discu

headlines: G tra

A 7.4 ma
the coast
U.S. Geolo

The epic
southwest

Reuters re
felt as fa
were no i
damage .

headlines: fixy bike festiva

GUATEMALA
Survey says tha
hit off the Pacifi
the capital and
as Mexico City

The U.S. Pacific
said there was a
tsunami, within
epicenter, but th
an immediate w
region.

The magnitude-
deep, was cente
Champerico.

People fled buil
Mexico City and

ktla.com

local • news • classifieds

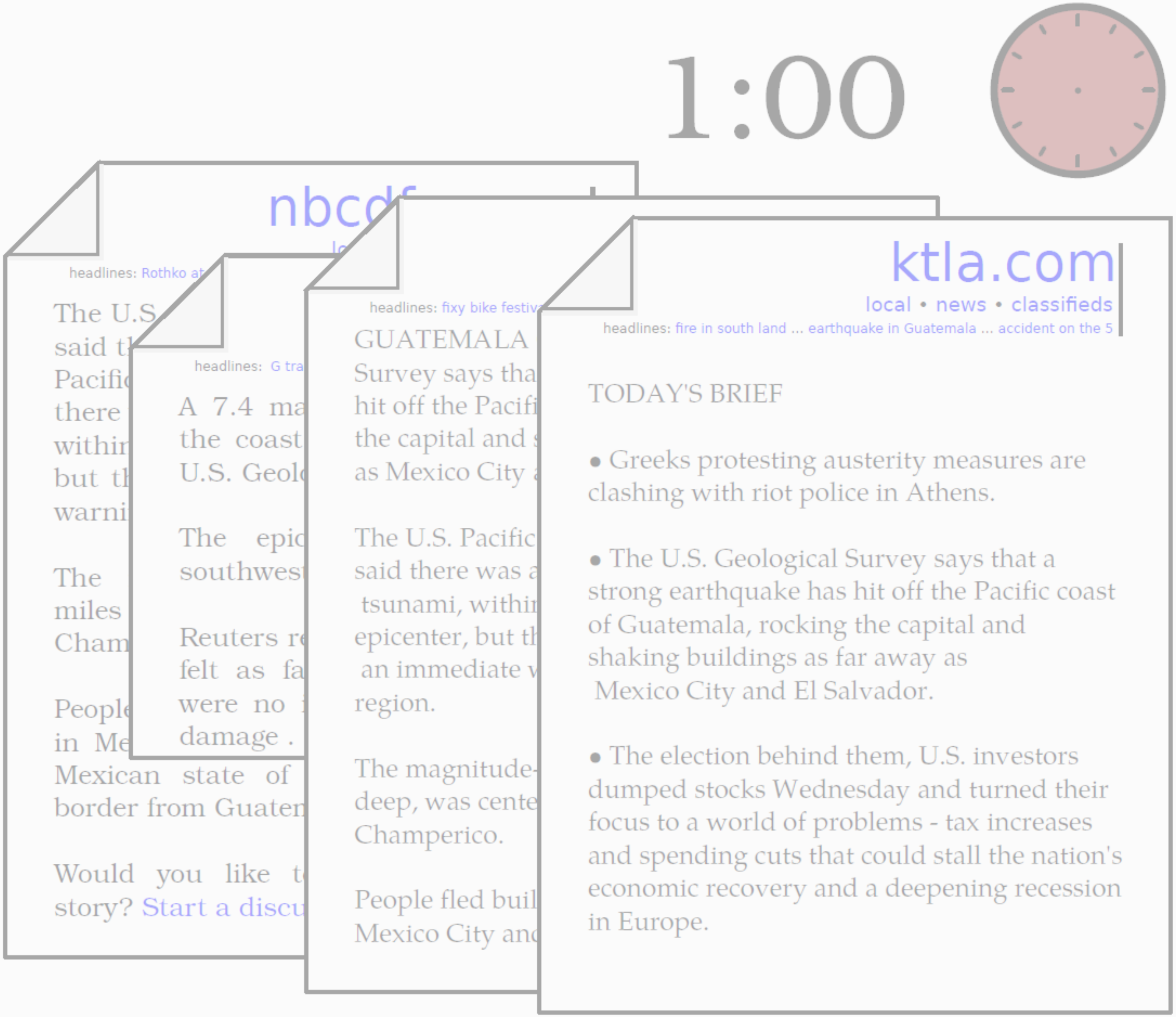
headlines: fire in south land ... earthquake in Guatemala ... accident on the 5

TODAY'S BRIEF

- Greeks protesting austerity measures are clashing with riot police in Athens.

- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.

- The election behind them, U.S. investors dumped stocks Wednesday and turned their focus to a world of problems - tax increases and spending cuts that could stall the nation's economic recovery and a deepening recession in Europe.



1:00

nbcdf

headlines: Rothko at

The U.S
said t
Pacifi
there
withi
but t
warni

The
miles
Cham

People
in Me
Mexican state of
border from Guaten

Would you like t
story? Start a discu

headlines: G tra

A 7.4 ma
the coast
U.S. Geolo

The epic
southwest

Reuters re
felt as fa
were no
damage .

headlines: fixy bike festiv

GUATEMALA
Survey says tha
hit off the Pacifi
the capital and
as Mexico City

The U.S. Pacific
said there was a
tsunami, withi
epicenter, but th
an immediate w
region.

The magnitude-
deep, was cente
Champerico.

People fled buil
Mexico City and

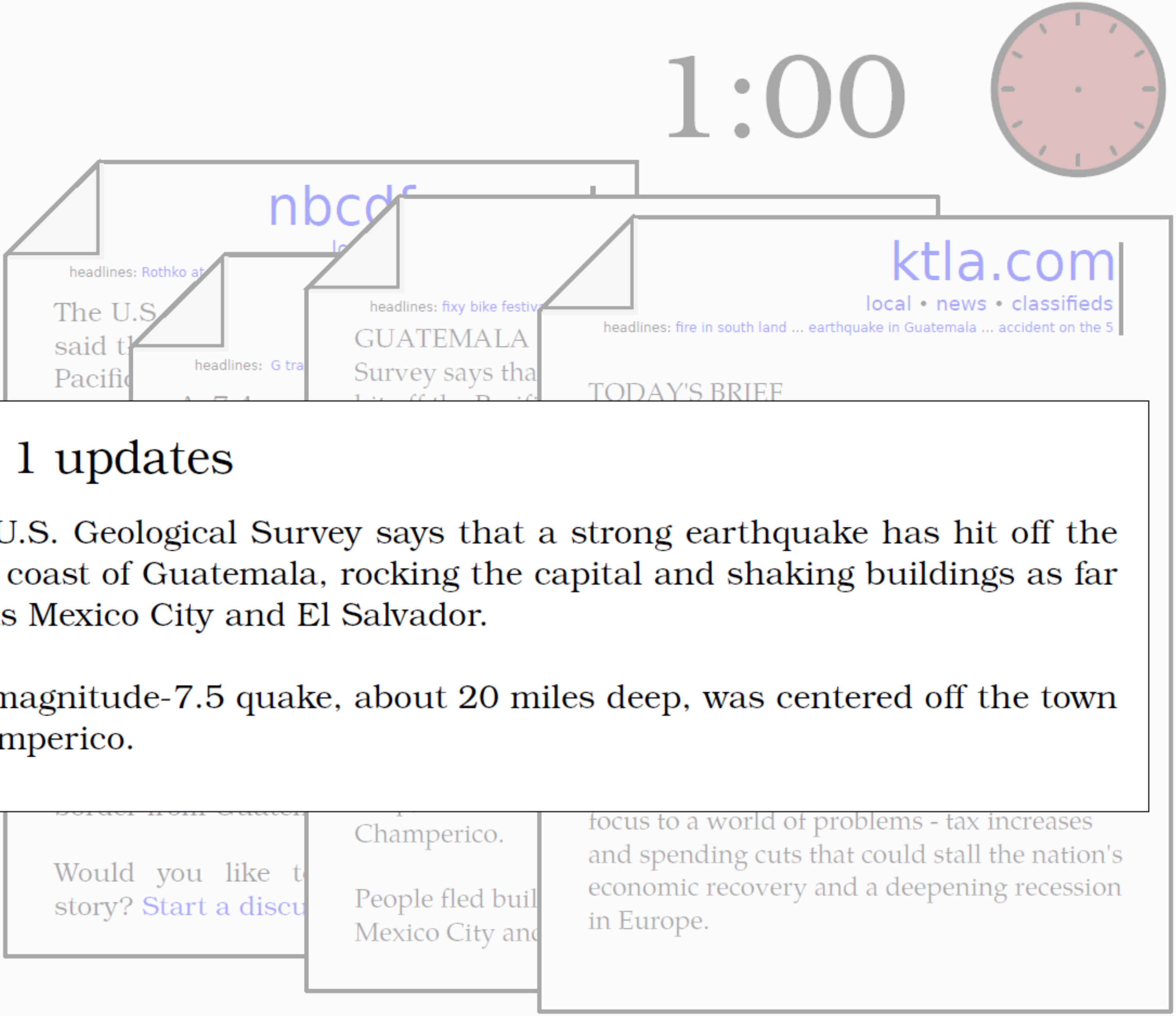
ktla.com

local • news • classifieds

headlines: fire in south land ... earthquake in Guatemala ... accident on the 5

TODAY'S BRIEF

- Greeks protesting austerity measures are clashing with riot police in Athens.
- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.
- The election behind them, U.S. investors dumped stocks Wednesday and turned their focus to a world of problems - tax increases and spending cuts that could stall the nation's economic recovery and a deepening recession in Europe.


1:00
nbcdf
headlines: Rothko at
The U.S
said t
Pacific
border from Guatem
Would you like t
story? Start a discu
headlines: G tra
A 7.4
headlines: fixy bike festiv
GUATEMALA
Survey says tha
Champerico.
People fled buil
Mexico City and
ktla.com
local • news • classifieds
headlines: fire in south land ... earthquake in Guatemala ... accident on the 5
TODAY'S BRIEF
focus to a world of problems - tax increases
and spending cuts that could stall the nation's
economic recovery and a deepening recession
in Europe.

hour 1 updates

- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.
- The magnitude-7.5 quake, about 20 miles deep, was centered off the town of Champerico.

Temporal Summarization Approach

At time **t**:

1. Predict salience for input sentences
   - Disaster-specific features for predicting salience
2. Remove redundant sentences
3. Cluster and select exemplar sentences for **t**
   - Incorporate salience prediction as a prior

Kedzie & al, Bloomberg Social Good Workshop, KDD 2014
Kedzie & al, ACL 2015

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

- generic news corpus (10 years AP and NY Times articles)
- domain specific corpus (disaster related Wikipedia articles)

A language model scores sentences by how typical they are of the language – higher scores mean more fluent

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

- generic news corpus (10 years AP and NY Times articles)
- domain specific corpus (disaster related Wikipedia articles)

A domain specific language model scores sentences by how typical they are of the disaster type

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

- generic news corpus (10 years AP and NY Times articles)
- domain specific corpus (disaster related Wikipedia articles)

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

Medium Salience

People streamed out of homes, schools and office buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

Geographic Features

- tag input with Named-Entity tagger
- get coordinates for locations and mean distance to event

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

Medium Salience

People streamed out of homes, schools and office buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

Geographic Features

- tag input with Named-Entity tagger
- get coordinates for locations and mean distance to event

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

Medium Salience

People streamed out of homes, schools and office buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

Geographic Features

Semantics

- number of event type synonyms, hypernyms, and hyponyms

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

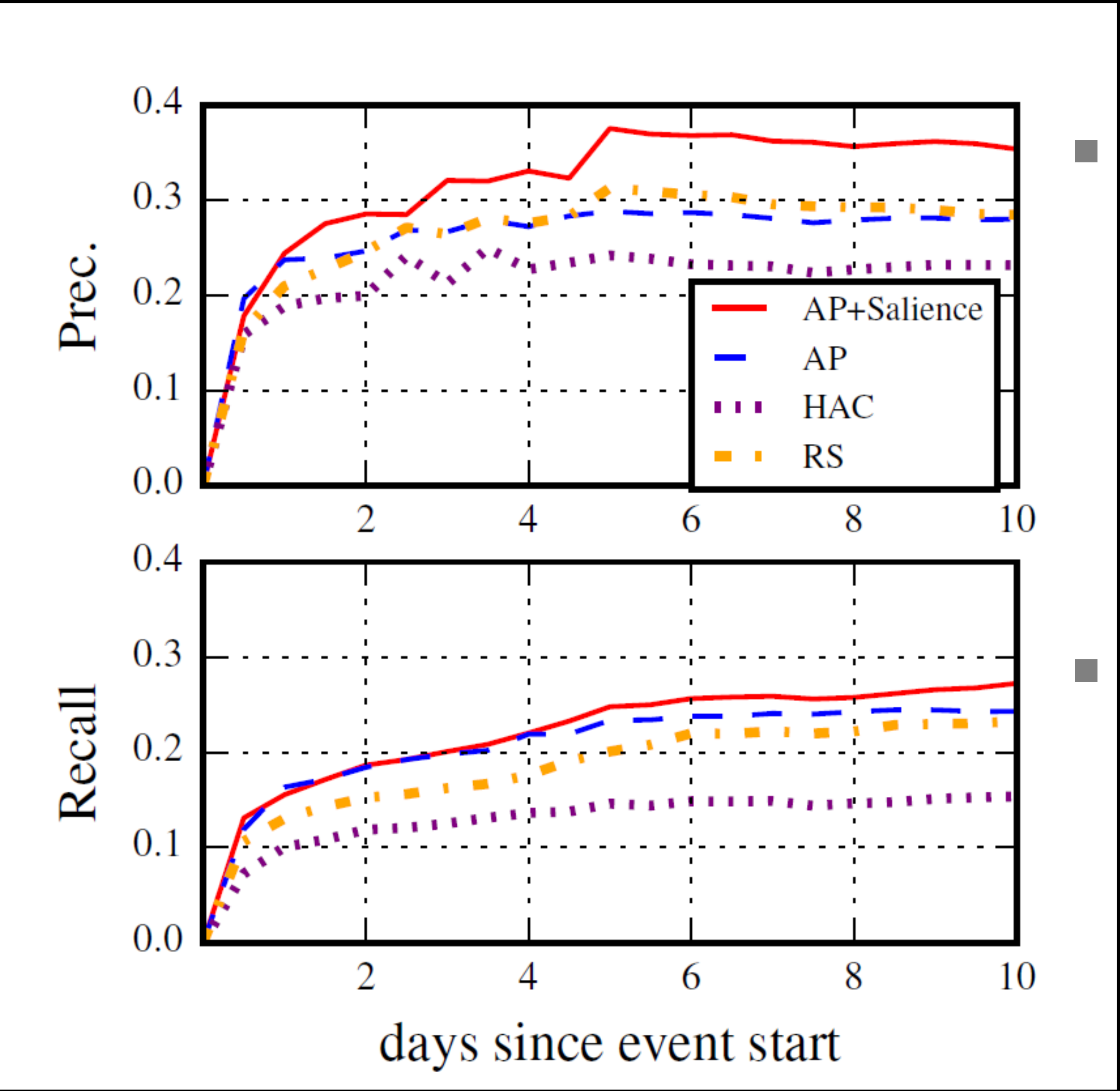
Medium Salience

People streamed out of homes, schools and office buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

What Have We Learned?



- Salience predictions lead to high precision quickly

- Salience predictions allow us to more quickly recover more information

Next Time

- Neural Net approaches to summarization
- HW4 released