

# *Text Summarization*

# *Announcements*

- HW4 out today.
- Final exam: Dec. 21<sup>st</sup>, 1:10-4pm
- Midterm curve: out tomorrow
- DSI/CS Distinguished Lecture: Yann LeCun, Monday 11/20 11AM Davis Auditorium
- AlphaGo documentary free screening. 5:30pm, Tuesday November 21, Roone Arledge Cinema, Lerner Hall. Register:

<https://www.eventbrite.com/e/movie-screening-alphago-tickets-39963928185>

# *Today*

- Multi-document abstractive summarization
- Headline generation (the basis for your homework)

# *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/>



Ani Nenkova

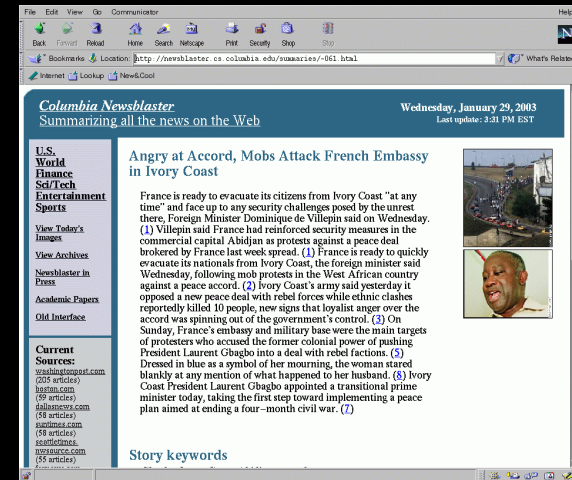


David Elson

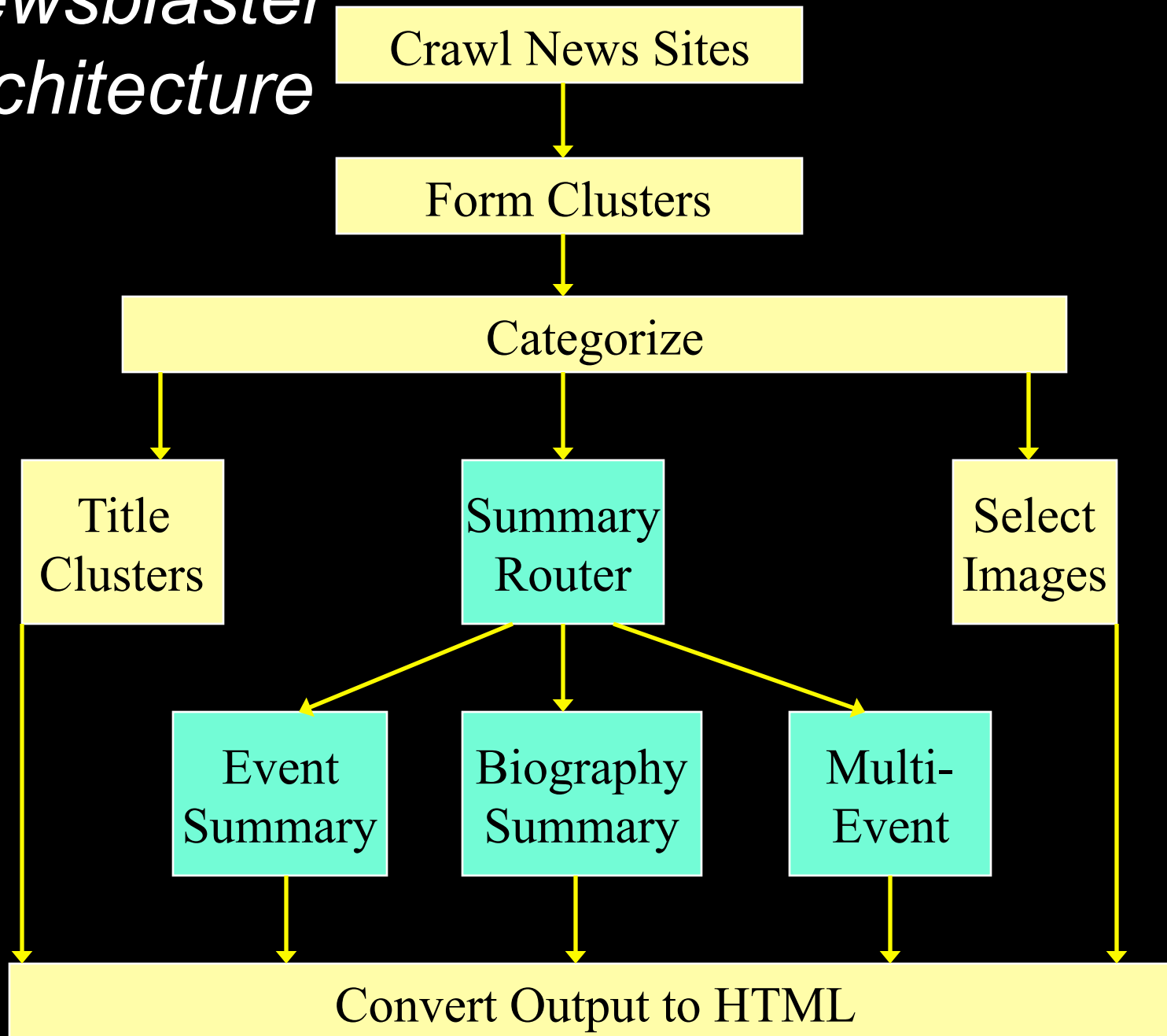
# Newsblaster

<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





# 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:

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

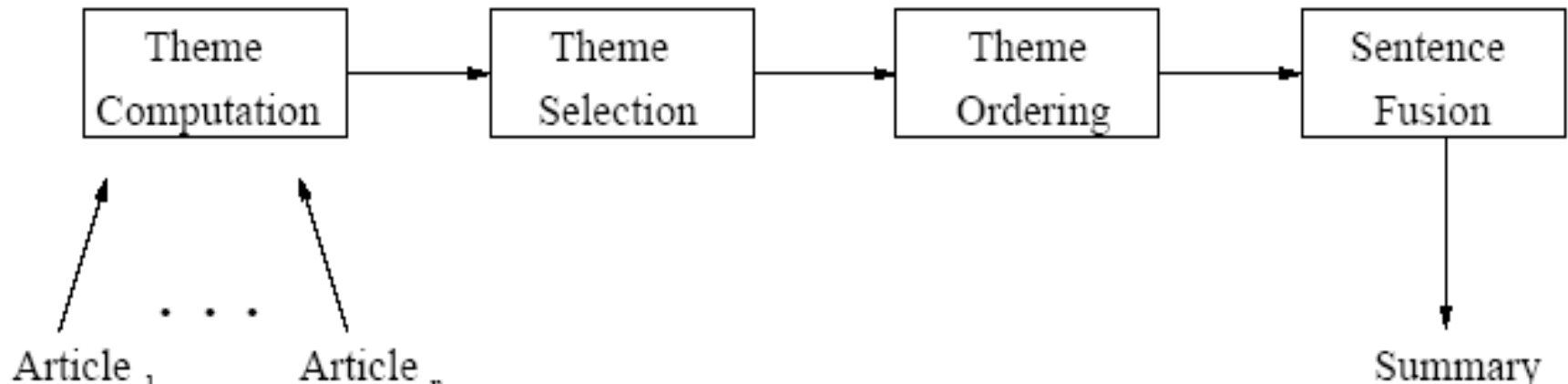
### [Tembec loses \\$51.5 million in fourth quarter; lumber joint venture lagging](#) (Finance, 8 articles)

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.

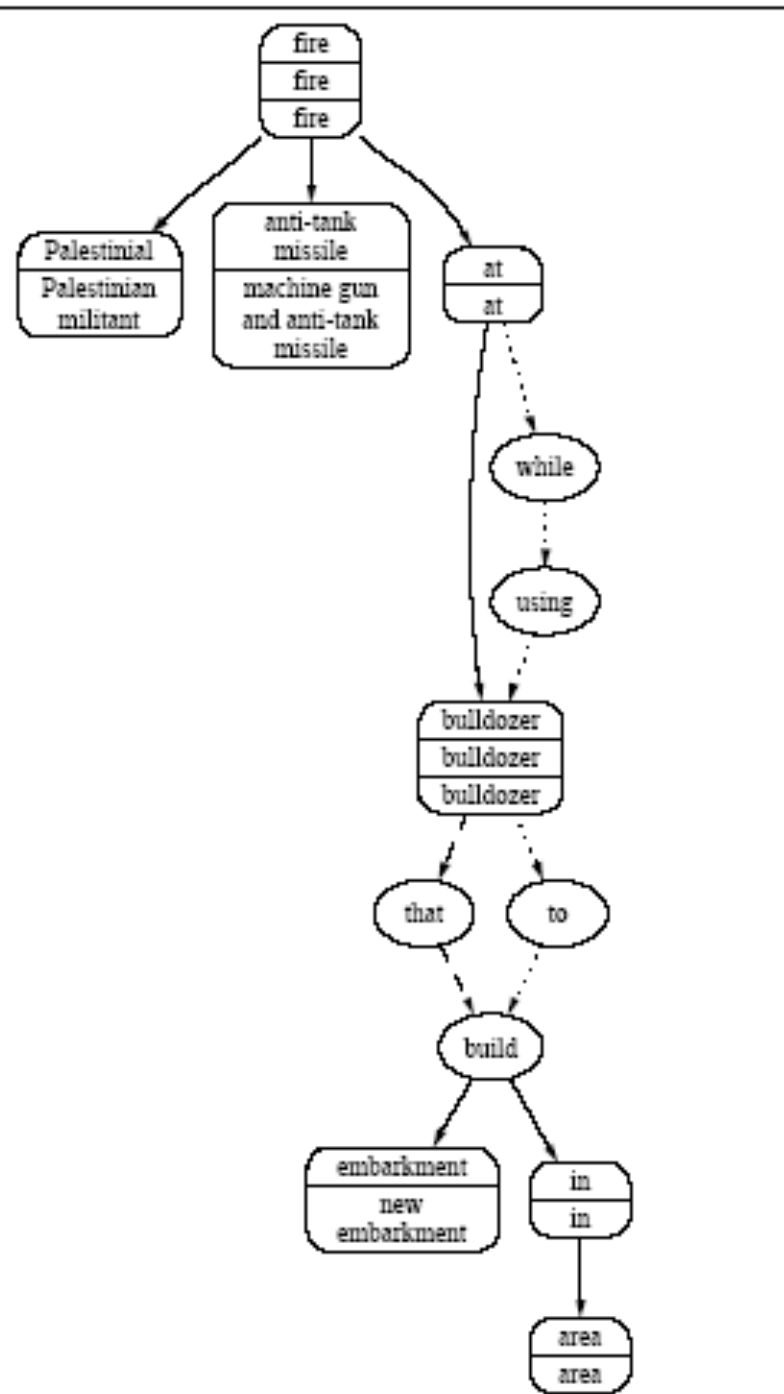
**fusion sentence:** Palestinians fired an anti-tank missile at a bulldozer.

# *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*

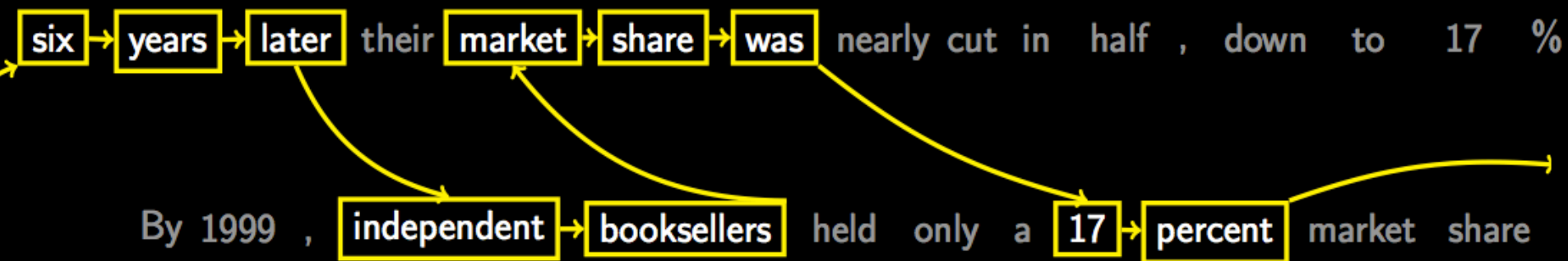
- Common information identification
  - Alignment of constituents in parsed theme sentences:  
*only some subtrees match*
  - Bottom-up local multi-sequence alignment
  - Similarity depends on
    - ◆ Word/paraphrase similarity
    - ◆ Tree structure similarity
- Fusion lattice computation
  - Choose 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



# Sentence Fusion – Structured Prediction



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



## 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:


[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](#)]

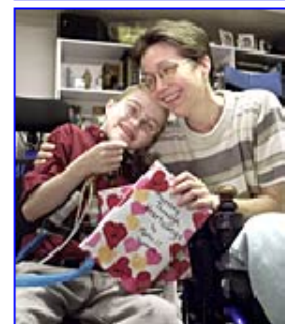
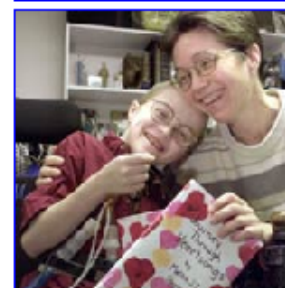
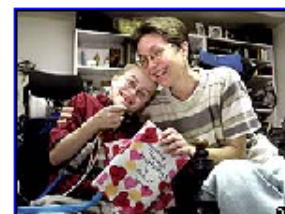
### Event tracking:

- [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?
- 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

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:

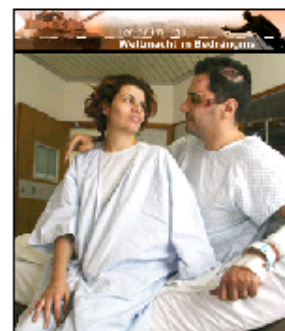
## U.S. pledges to help Saudi war on terror after weekend attacks

Summary from multiple countries, from articles in multiple languages

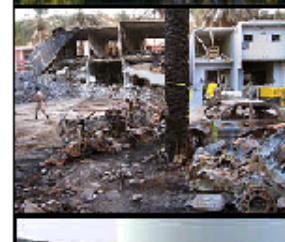
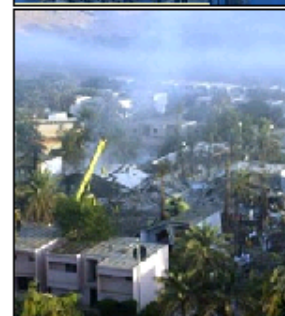
Saudi security officials are deploying thousands of troops to the city of Mecca because of concerns that terrorists may be planning new attacks during the Muslim holy month of Ramadan, Saudi government sources told CNN. U.S. to close embassy in Sudan, officials say The attack occurred a day after the 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](#)]



**TOP500**  
Die größten deutschen Unternehmen!  
Jetzt auch als Download!



# 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

### 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

# AKTUELLES

- HOME
- WELT AM SONNTAG
- AKTUELL
- POLITIK
- WIRTSCHAFT
- FINANZEN
- IMMOBILIEN
- SPORT
- VERMISCHTES
- KULTUR
- MEDIEN
- WISSENSCHAFT
- FORUM
- MAGAZIN
- HAMBURG
- BERLIN
- BREMEN
- REISEWELT
- LITERARISCHE WELT
- AUTO & BOOT
- KARRIEREWELT
- BUSINESS EXPLORER
- ABONNEMENT
- ANMELDUNG
- ARCHIV
- IMPRESSUM
- KONTAKT
- MEDIAWELT
- TV-PROGRAMM

Montag, 17. November 2003 Berlin, 04:19 Uhr **DIE WELT**

## 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



### news TICKER Themen heute

- 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



### dax INTRADAY



# *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

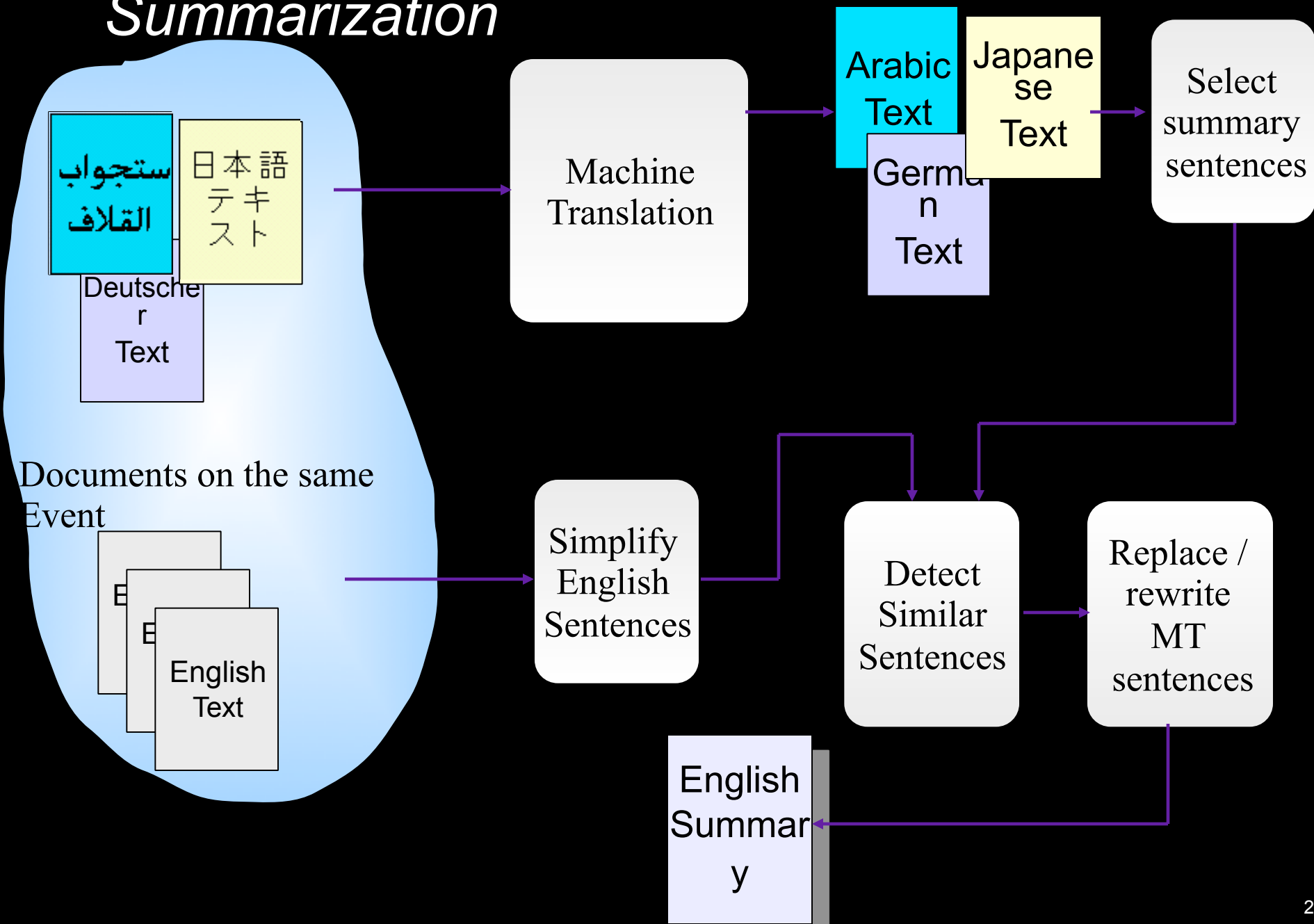
The Iraqi being stationed US military shot on the 20th, the same day to the allied forces detention facility which is in アブグレ アブグレイブ hdad west approximately 20 kilometers, *mortar 12 shot* and you were packed, *22 Iraqi human prisoners died*, it announced that nearly 100 people were injured.

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



# 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

↓


name	Nizar Hamdoon
role	representative
country	Iraq ( <i>arg1</i> )
organization	United Nations ( <i>arg2</i> )

↓

Iraqi United Nations representative Nizar Hamdoon



some likely problems with this approach to  
NPs?



**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](https://PollEv.com/app)

# *Current work*

- <http://datascience.columbia.edu/14-million-dollar-computer-system-to-translate-and-summarize-documents>

# *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)*

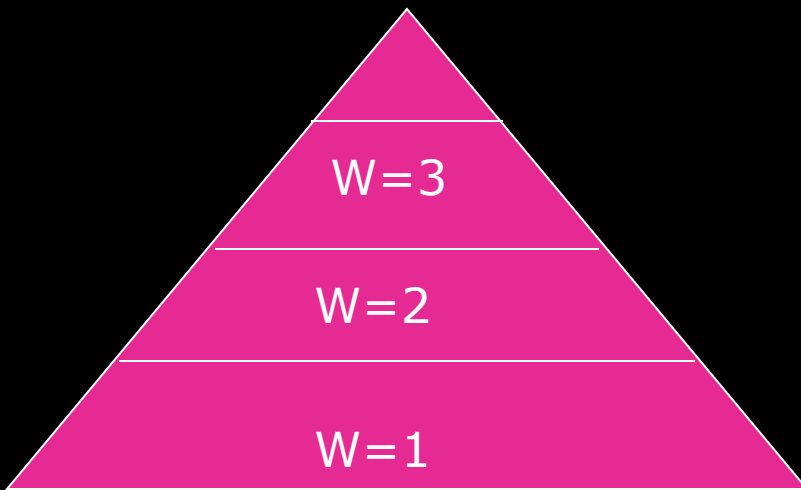
- A. The cause of the fire was unknown.
- B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.
- C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in 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*

$$\text{SCORE} = D/\text{MAX}$$

**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 browser'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

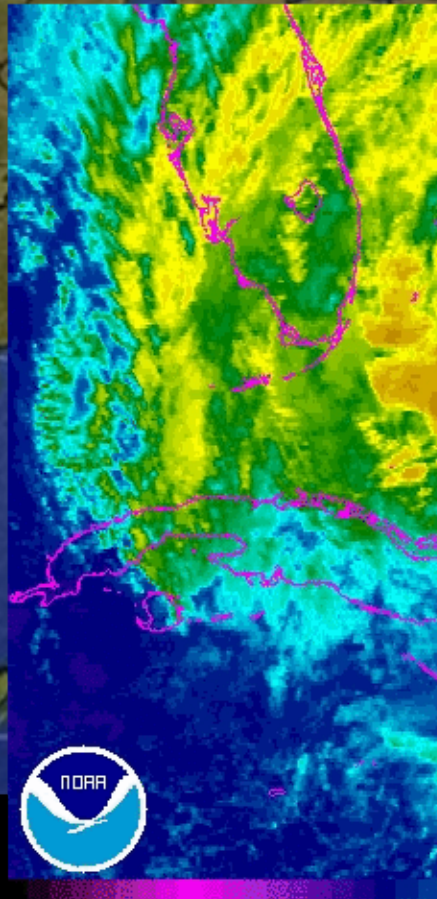
# Questions (from Sparck Jones)

- 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



## HURRICANE SANDY

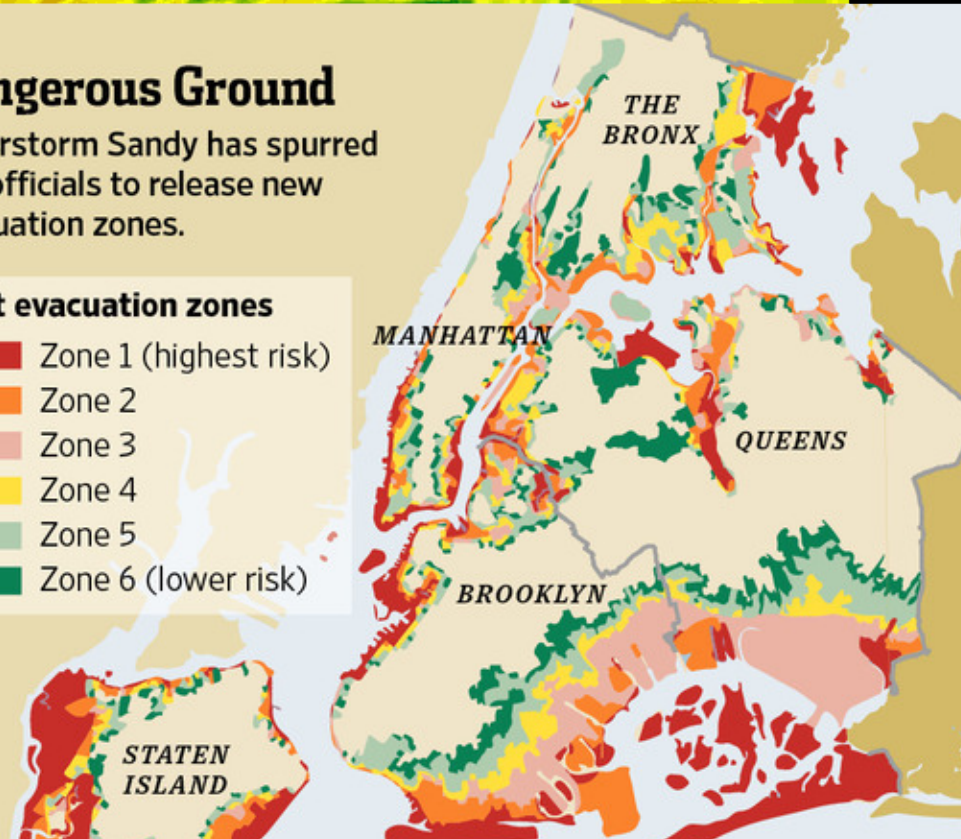


### Dangerous Ground

Superstorm Sandy has spurred city officials to release new evacuation zones.

#### Draft evacuation zones

- Zone 1 (highest risk)
- Zone 2
- Zone 3
- Zone 4
- Zone 5
- Zone 6 (lower risk)





# *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



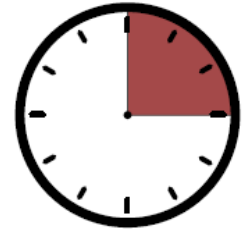
Public Transit Outage

# 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

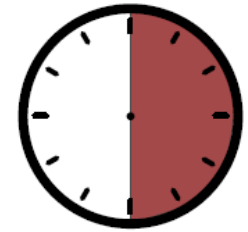
The U.S. Pacific Tsunami Warning Center said there was a possibility of a local U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.

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 of the 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  
withi  
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



headlines: Rothko at

The U.S. said the Pacific there within but the warni

headlines: G tra

A 7.4 ma the coast U.S. Geol

The epic southwes

Reuters re felt as fa were no damage .

Mexican state of border from Guater

Would you like to story? [Start a disc](#)

nbcdf

kgw.com  
local • news • classifieds

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

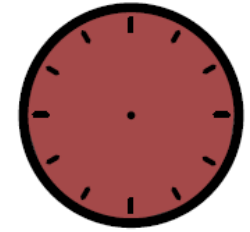
GUATEMALA CITY -- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.

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



headlines: Rothko at ...

The U.S. said the Pacific there within but the warni

The miles Cham

People in Me

Mexican state of border from Guater

Would you like to story? [Start a disc](#)

nbcdf

headlines: fixy bike festi

GUATEMALA

Survey says that hit off the Pacific the capital and s as Mexico City a

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

The magnitude-deep, was cente Champerico.

People fled buil Mexico City and

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

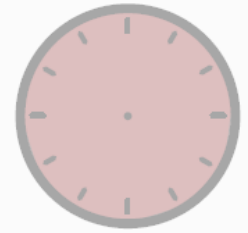
ktla.com

local • news • classifieds

### 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



headlines: Rothko at ...

The U.S. said the Pacific there within but the warni

The miles Cham

People in Me

Mexican state of border from Guater

Would you like to story? [Start a disc](#)

headlines: fixy bike festi

GUATEMALA

Survey says that hit off the Pacific the capital and s as Mexico City a

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

The magnitude-deep, was cente Champerico.

People fled buil Mexico City and

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

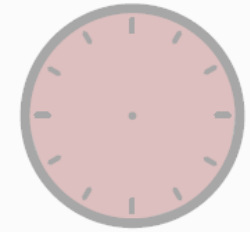
ktla.com

local • news • classifieds

### 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



## 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.

Would you like to  
story? [Start a disc](#)

Champerico.  
People fled build  
Mexico City and

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.



# *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

# *Predicting Saliency: 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.

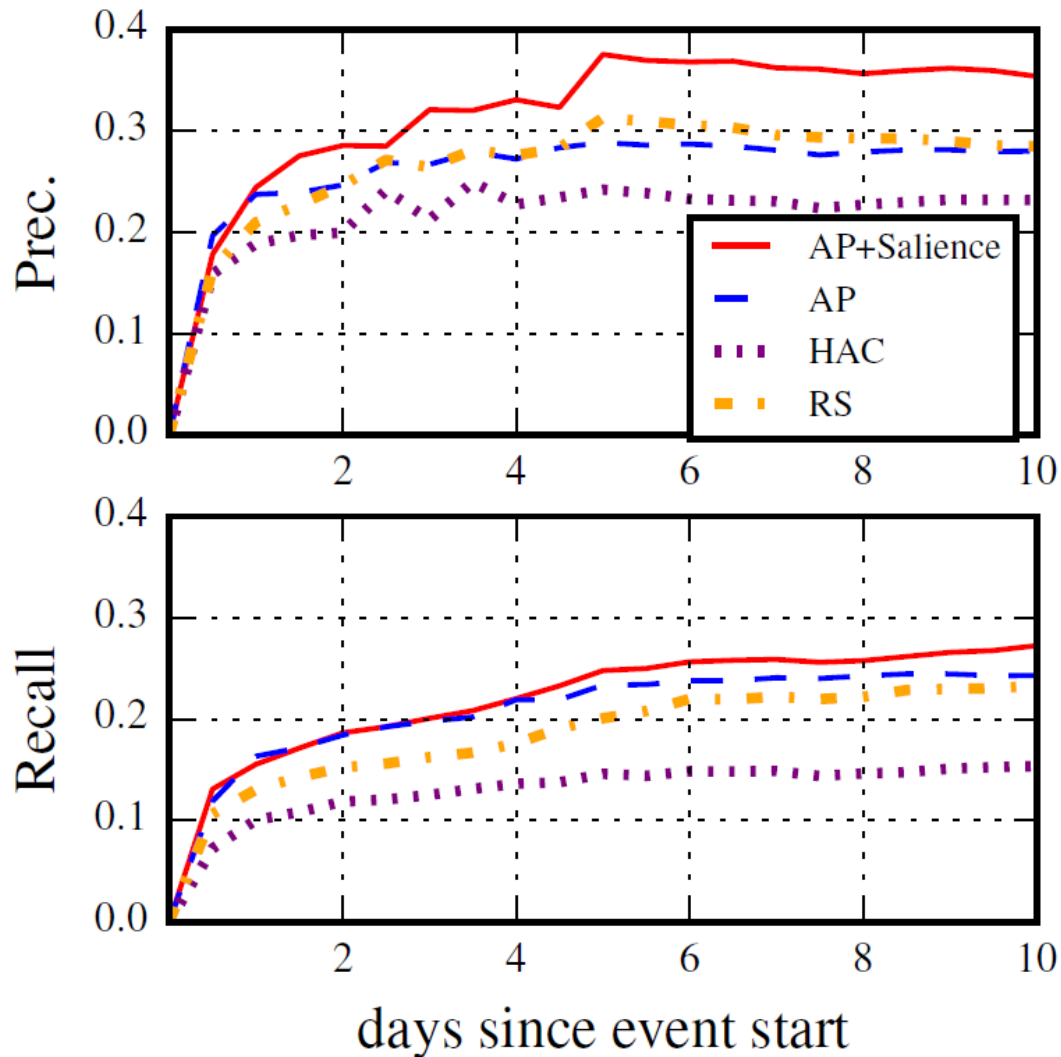
## **Medium Salience**

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?



- Saliency predictions lead to high precision quickly

- Saliency predictions allow us to more quickly recover more information

# *A Neural Model*

- A Neural Attention Model for Abstraction Summarization Alexander Rush, Sumit Chopra, Jason Weston
- Basis for HW4

# *Task: Headline Generation*

- Input: first sentence of a news article
- Output: headline
- **Attention Based Summarizer (ABS):**  
Performs abstractive summarization
  - Paraphrasing, compression

# *Training Data*

- 4 million article/headline pairs from Gigaword
  - A detained iranian-american academic accused of acting out against national security has been released from a tehran prison yesterday after a hefty bail was posted, a top judiciary official said Tuesday.  
Detained iranian-american academic released from prison after hefty bail.
  - Ministers from the european union and its mediterranean neighbors gathered here under heavy security on Monday for an unprecedented conference on economic and political cooperation.  
European mediterranean ministers gather for landmark conference by julie bradford.

# *System Output*

- russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism.
- russia calls for joint front against terrorism.



# Problem Framework

- Input:  $\mathbf{x}_1 \dots \mathbf{x}_m$
- Output:  $\mathbf{y}$  of length  $n < m$ 
  - Output length  $n$  fixed and known by system
- Fixed vocabulary  $V$  of size  $|V|$  and both input and output come from  $V$
- Scoring function  $S(\mathbf{x}, \mathbf{y}) = \log P(\mathbf{y}|\mathbf{x}; \Theta)$  equivalent to:

$$\log p(\mathbf{y}|\mathbf{x}; \theta) \approx \sum_{i=0}^{N-1} \log p(\mathbf{y}_{i+1}|\mathbf{x}, \mathbf{y}_c; \theta)$$

- Where  $\mathbf{y}_c$  is previous context and  $c$  limited using Markov assumption

# *Main Focus*

- Modeling the local conditional distribution:
  - $P(y_{i+1} | x_i, y_c; \Theta)$
- Neural network is a seq-to-seq model
  - Encoder: a conditional summarization model
  - Decoder: Neural language probabilistic model

# Neural Language Model

- Standard Neural Network Language Model (NNLM) (Banks et al 2003)

$$\begin{aligned} p(\mathbf{y}_{i+1} | \mathbf{y}_c, \mathbf{x}; \theta) &\propto \exp(\mathbf{V}\mathbf{h} + \mathbf{W}\text{enc}(\mathbf{x}, \mathbf{y}_c)), \\ \tilde{\mathbf{y}}_c &= [\mathbf{E}\mathbf{y}_{i-C+1}, \dots, \mathbf{E}\mathbf{y}_i], \\ \mathbf{h} &= \tanh(\mathbf{U}\tilde{\mathbf{y}}_c). \end{aligned}$$

- Parameters:  $\theta = (\mathbf{E}, \mathbf{U}, \mathbf{V}, \mathbf{W})$
- E is a word embedding matrix
- U, V, W are weight matrices

# *Experiments with 3 encoders*

## 1. Bag of word encoder

- Bag of words of input sentence embedded down to  $H$ , size of hidden layer
- Ignores order and context in input
- Captures relative importance of words

# *Experiments with 3 Encoders*

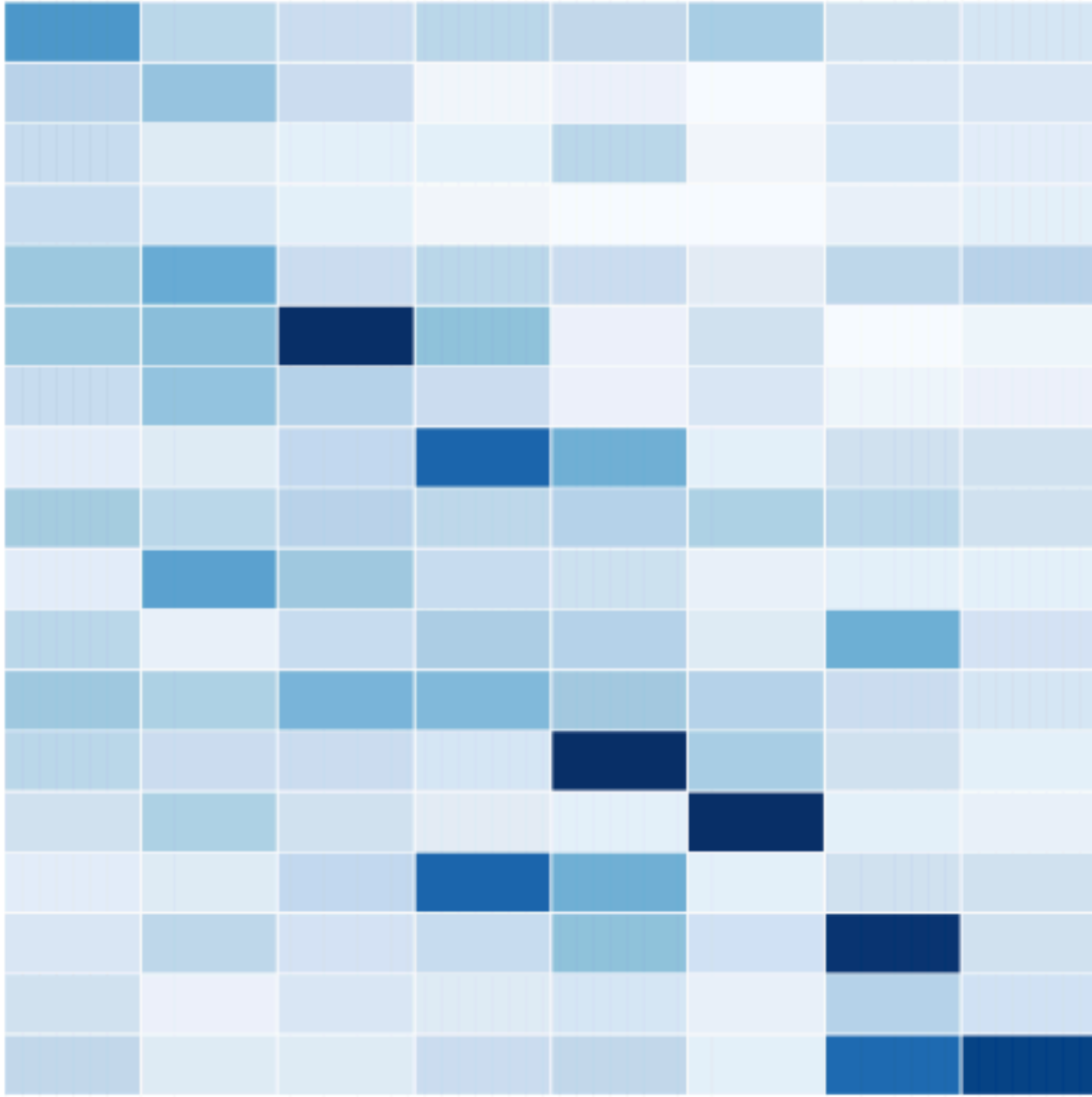
## 2. Convolutional encoder

- Allows for local interactions in the input
  - CNNs known to be good at capturing patterns such as n-grams
- Does not require encoding input context  $y_c$
- But it must produce a single representation for the entire sentence

# *Experiments with 3 Encoders*

- Attention-based encoder
  - Constructs a representation using generation context ( $y_c$ )
  - Replaces uniform distribution of bag of words with a learned soft alignment between input and summary
    - Used to weight the smoothed input when construction the representation

<S>    russia    calls    for    joint    front    against    terrorism



<s>  
 russian  
 defense  
 minister  
 ivanov  
 called  
 sunday  
 for  
 the  
 creation  
 of  
 a  
 joint  
 front  
 for  
 combating  
 global  
 terrorism

# *Interesting extensions*

- Uses a beam search decoder
  - Maintains full vocabulary
  - Limits to  $K$  potential hypotheses at each position of the summary
- Add features to promote using words of input (extractive features)
  - Combine the local conditional probability with indicator features for  $n$ -grams from input



# *Experiments*

- Rouge-1, -2 and -L as metrics
- On headlines:
  - 17 point jump in Rouge-1 from worst baseline (IR)
  - 11 point jump in Rouge 1 from compressive system
- On DUC-2004, 17 and 10 point jumps

# *Next*

- LSTM
- Another neural summarization approach: extractive document summarization