Text Summarization

Announcements

- Reading
  - Summarization: two neural net papers
  - Semantics: not responsible for C 13
- Class participation
- My office hours: please visit!
  - M 2:30-3:30
  - W 5-6

Today

- Summarization
  - Introduction
  - Extractive methods
  - Abstractive methods
  - Summarizing multiple documents: Updates on disaster
  - Evaluation

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

Extractive Summarization

- Input: single document or multiple documents

- Task: Select sentences that are salient/ representative of the input
  - Sentences strung together to produce output summary

- Summaries can be of varying lengths.

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

How do we determine what is salient?

- Much work has been unsupervised
- Later work uses machine learning to train a system
  - Requires large corpus of article/summary pairs

Methods for salience

- Term frequency*Inverse document frequency (TF*IDF)
- Log likelihood ratio
- Graph based methods

*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

An example

- New York Times front page article (Monday, 11/4): https://www.nytimes.com/2019/11/03/business/drunk-driving-breathalyzer.html
- What words occur in just about every news article every day (our background corpus)?
- What words are both frequent in this article and unique to this article?

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, Textrank
  - Erkan and Radev
  - Mihalcea

Topic Signature Words

- Uses the log ratio test to find words that are highly descriptive of the input
- Threshold to divide all words in the input into either descriptive or not
  - H1: the probability of a word in the input is the same as in the background
  - H2: 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

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

Log likelihood ratio

λ=L(H1)/L(H2)

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

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-squared

Graph-based methods

- Sentences vote for other sentences
  - Frequently occurring words link many sentences
- Input represented as highly connected 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 probability 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

A million Americans a year are arrested for drunken driving, and most stops begin the same way: flashing blue lights in the rearview mirror, then a battery of tests that might include standing on one foot or reciting the alphabet.

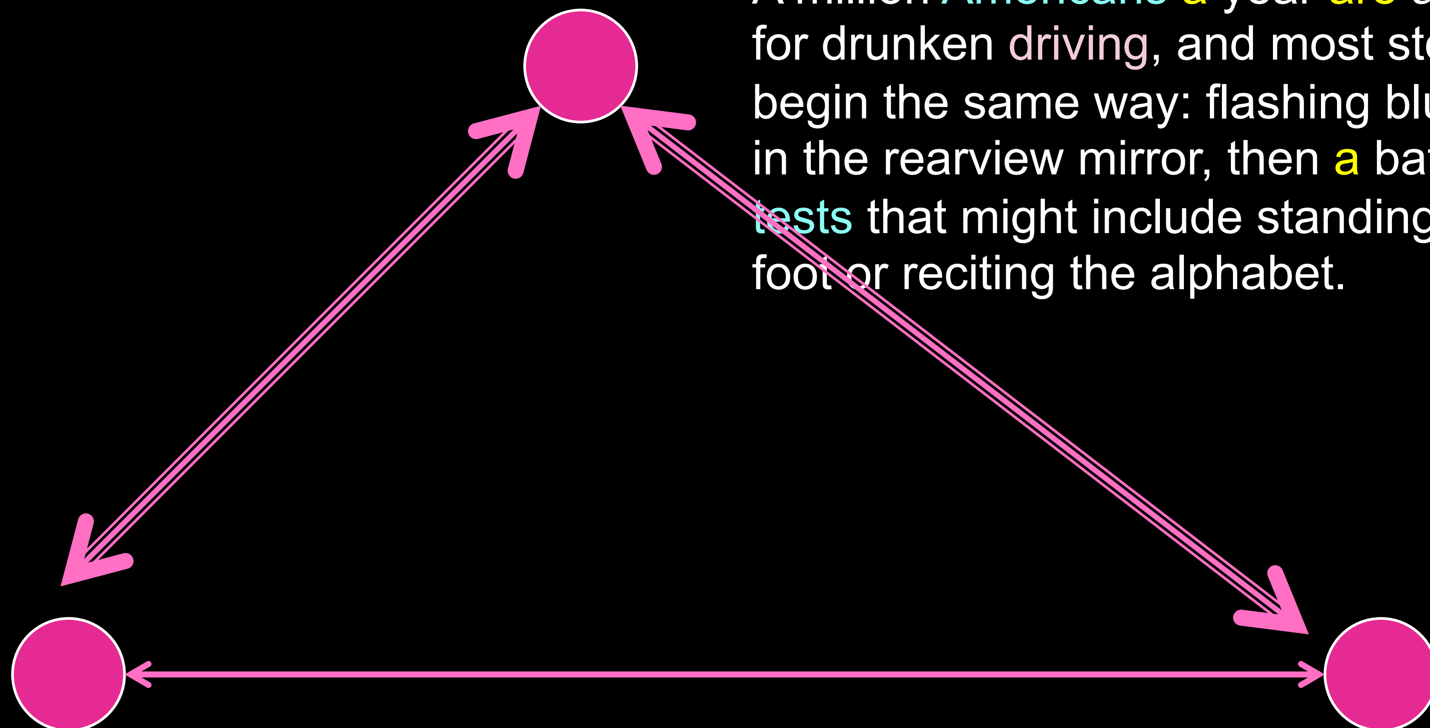
What matters most, though, happens next. By the side of the road or at the police station, the drivers blow into a miniature science lab that estimates the concentration of alcohol in their blood. If the level is 0.08 or higher, they are all but certain to be convicted of a crime.

But those tests — a bedrock of the criminal justice system — are often unreliable, a New York Times investigation found. The devices, found in virtually every police station in America, generate skewed results with alarming frequency, even though they are marketed as precise to the third decimal place.

A million Americans a year are arrested for drunken driving, and most stops begin the same way: flashing blue lights in the rearview mirror, then a battery of tests that might include standing on one foot or reciting the alphabet.

What matters most, though, happens next. By the side of the road or at the police station, the drivers blow into a miniature science lab that estimates the concentration of alcohol in their blood. If the level is 0.08 or higher, they are all but certain to be convicted of a crime.

But those tests — a bedrock of the criminal justice system — are often unreliable, a New York Times investigation found. The devices, found in virtually every police station in America, generate skewed results with alarming frequency, even though they are marketed as precise to the third decimal place.



A million Americans a year are arrested for drunken driving, and most stops begin the same way: flashing blue lights in the rearview mirror, then a battery of tests that might include standing on one foot or reciting the alphabet.

What matters most, though, happens next. By the side of the road or at the police station, the drivers blow into a miniature science lab that estimates the concentration of alcohol in their blood. If the level is 0.08 or higher, they are all but certain to be convicted of a crime.

But those tests — a bedrock of the criminal justice system — are often unreliable, a New York Times investigation found. The devices, found in virtually every police station in America, generate skewed results with alarming frequency, even though they are marketed as precise to the third decimal place.

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.

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

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

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

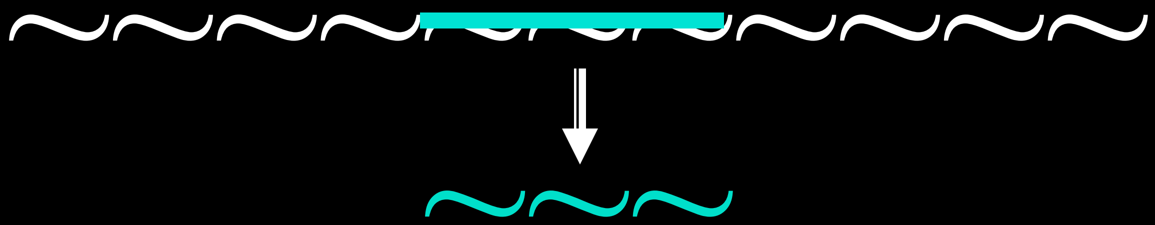
- (1) Sentence compression

- (2) Sentence fusion

Major Cut and Paste Operations

- (3) Syntactic Transformation

- (4) Lexical paraphrasing



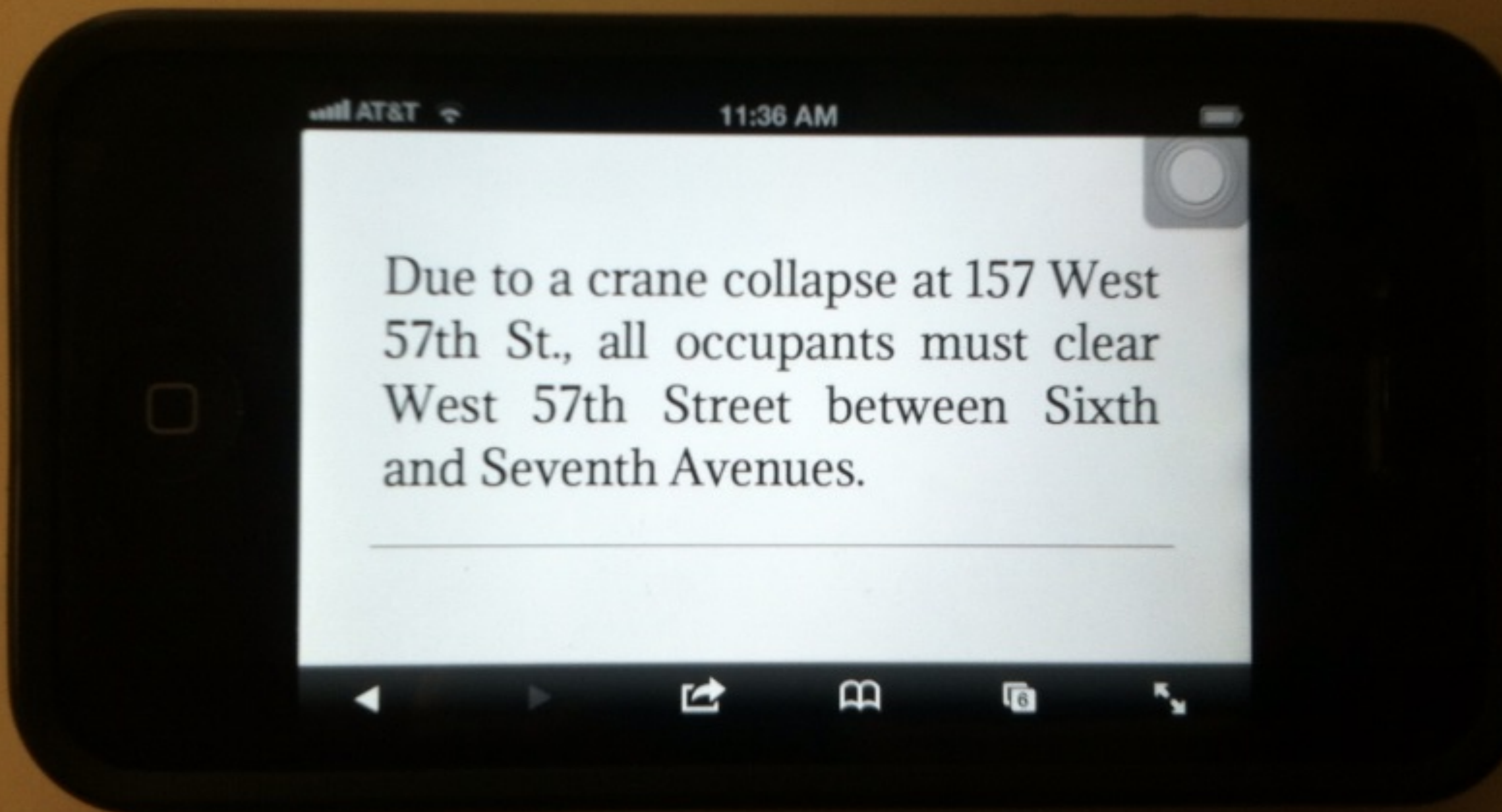
Summarization at Columbia

- 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



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

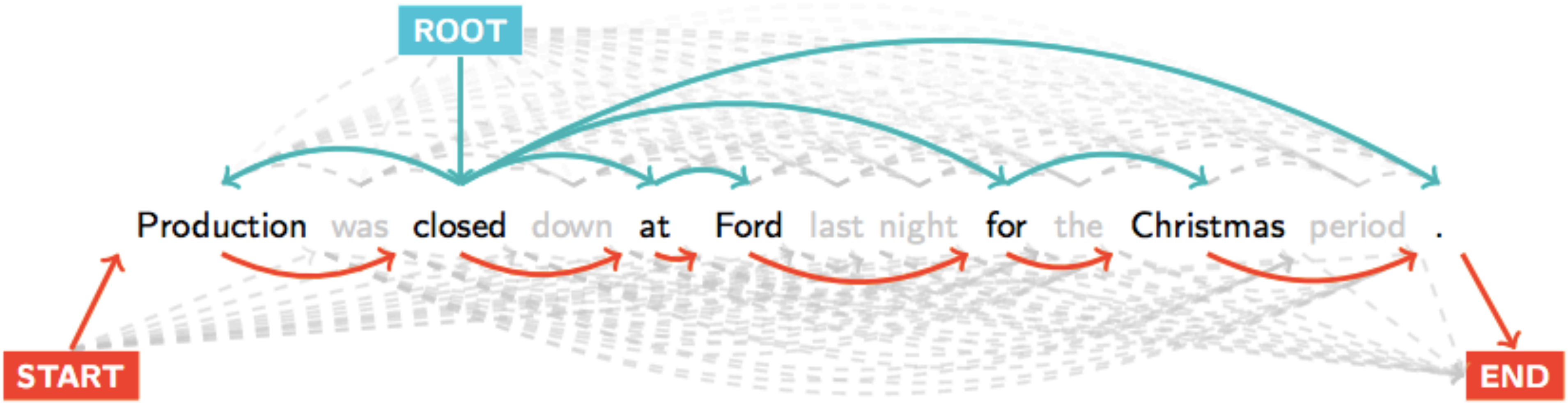
Joint inference over

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

Thadani & McKeown, CONLL 2013

structural factorizations

this work



Goal: recover tokens $\mathbf{x}$, n-gram sequence $\mathbf{y}$ and dependency structure $\mathbf{z}$

Slide from Thadani

Joint inference using Integer Linear Programming

- Objective that we want to maximize

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

$$C = \arg\max_{\mathbf{x},\mathbf{y},\mathbf{z}} \sum_i x_i \cdot \mathbf{w}_{tok}^\top \phi(t_i)$$ token

$$+ \sum_{i,j,k} y_{ijk} \cdot \mathbf{w}_{ngr}^\top \phi(\langle t_i, t_j, t_k \rangle)$$ ngram score

$$+ \sum_{i,j} z_{ij} \cdot \mathbf{w}_{dep}^\top \phi(\langle t_i, t_j \rangle)$$ dep score

- Token score: informativeness
- Ngram score: fluency
- Dependency score: fidelity

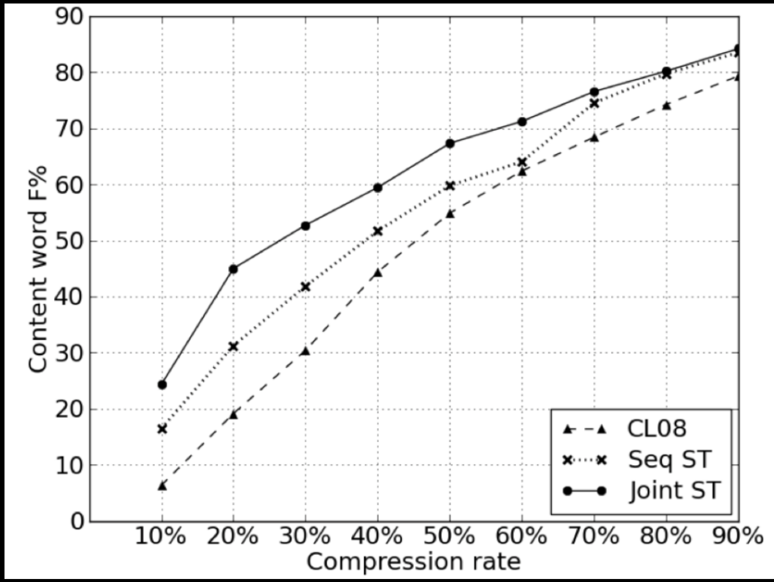
Compression

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

What Have We Learned?

Compression

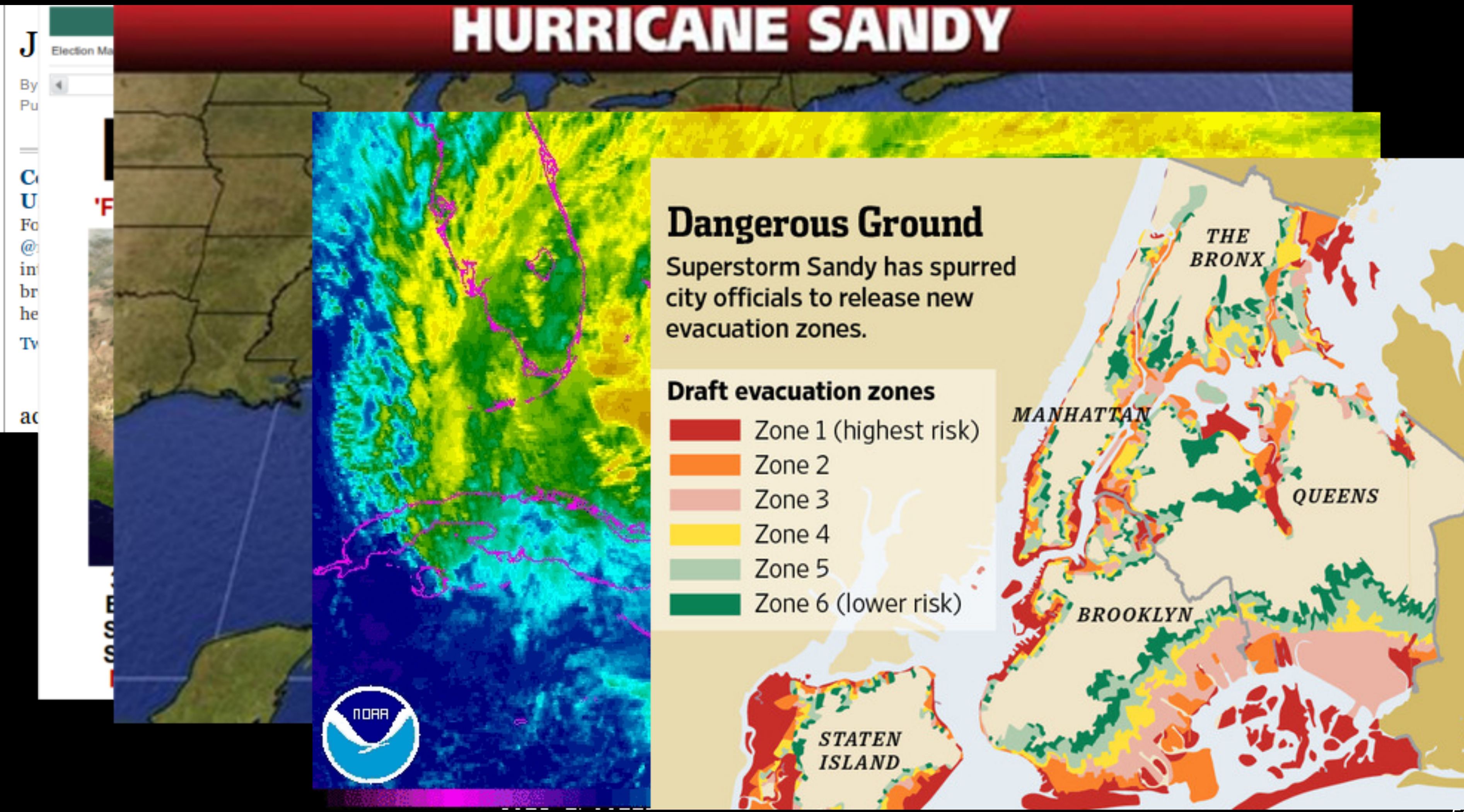
+5% n-gram recall for joint inference with dependency relations



Multi-Document Summarization A more common task

- 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

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

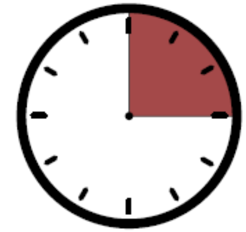
Manhattan Blackout

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

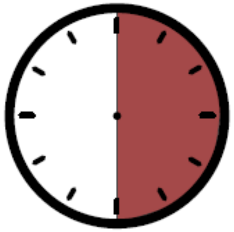
The U.S. Pacific Tsunami Warning Center said there was a possibility of a local U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.

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
Pacifi
there
withi
but t
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

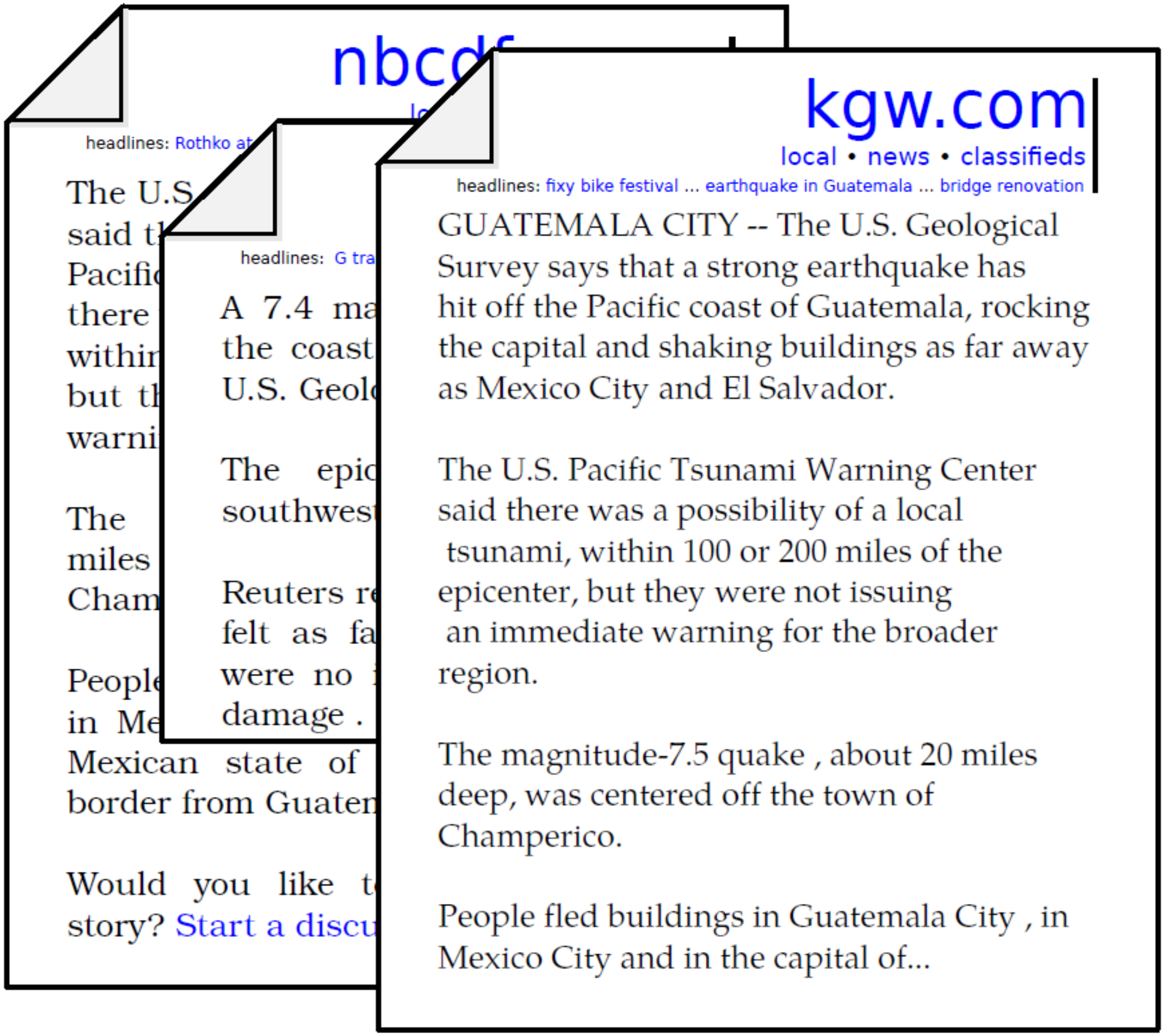
headlines: G train stuck forever ... weather on the 1's ... rats eat tourists

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



kgw.com

local • news • classifieds

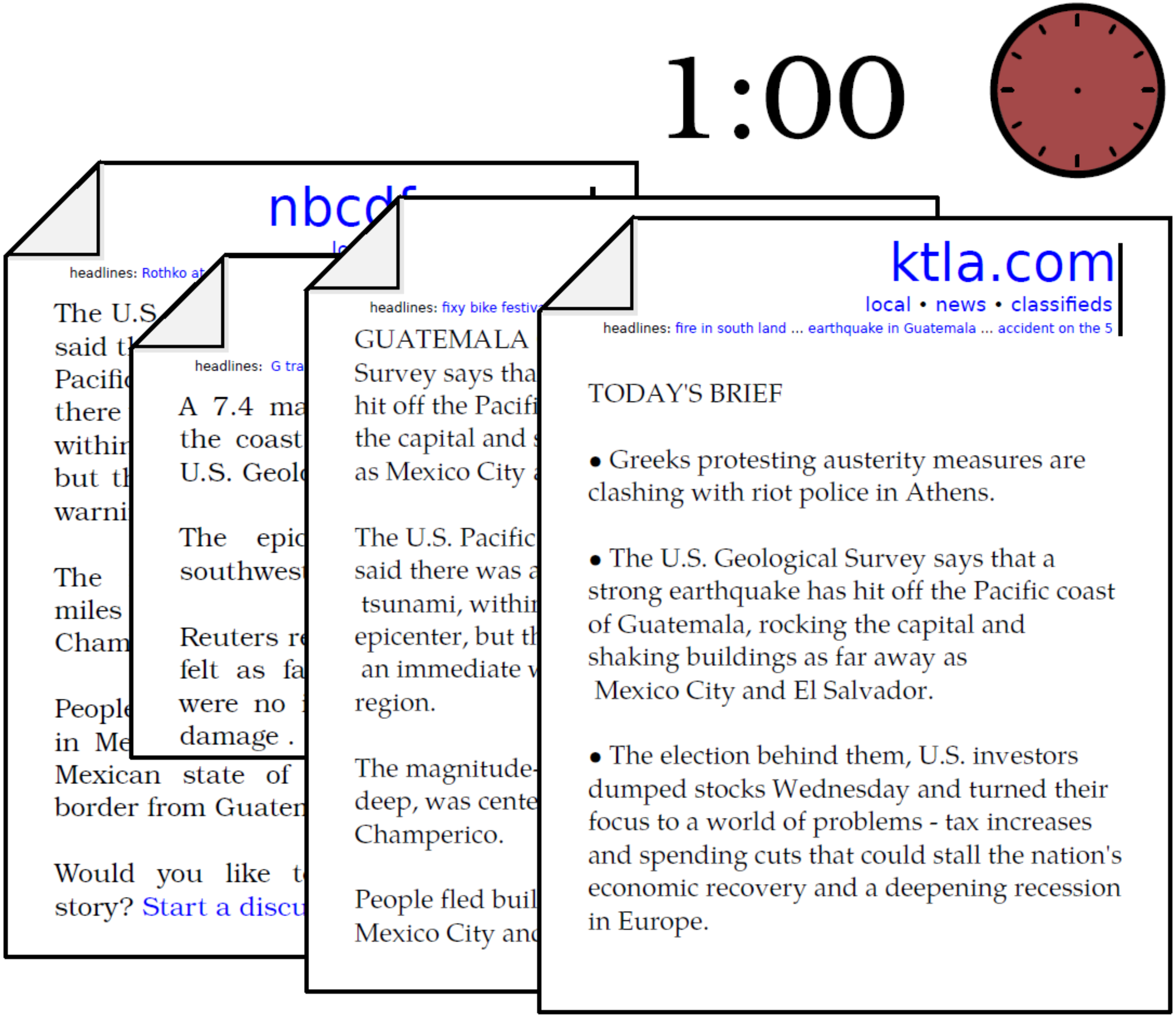
headlines: fixy bike festival ... earthquake in Guatemala ... bridge renovation

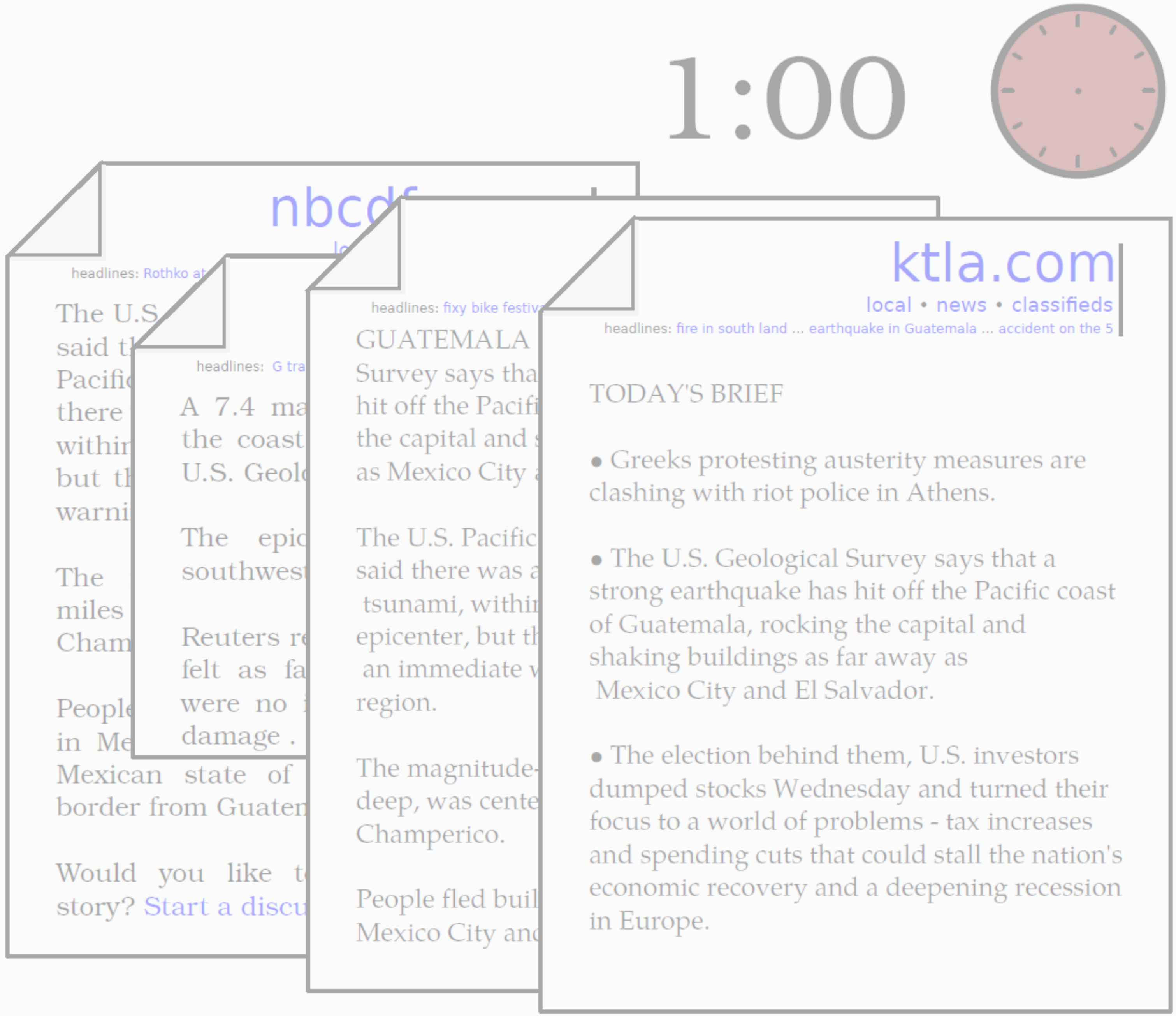
GUATEMALA CITY -- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.

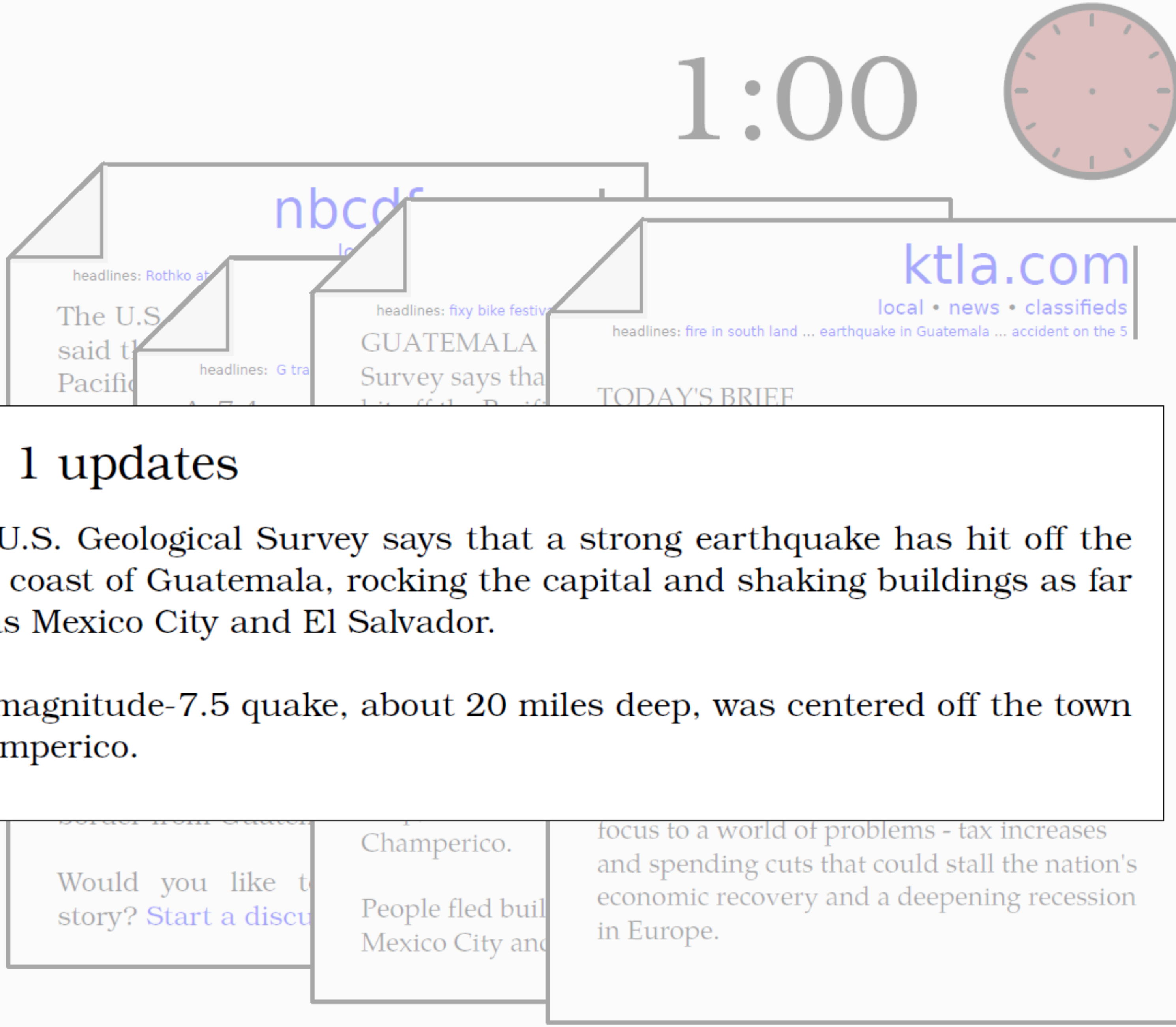
The U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.

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
headlines: Rothko at
headlines: G tra
headlines: fixy bike festiv
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
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.



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 oce 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 oce 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 oce 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.

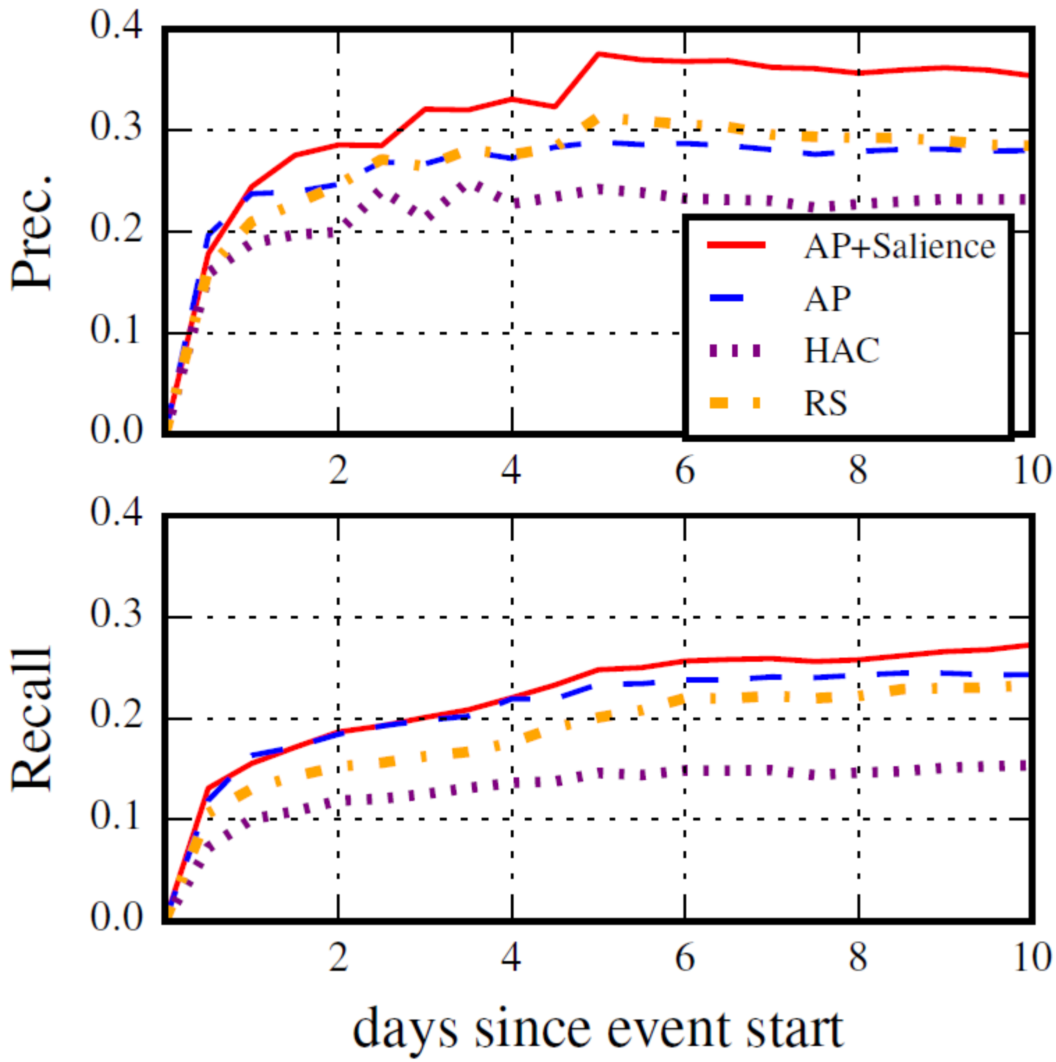
Medium Salience

People streamed out of homes, schools and oce 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

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)

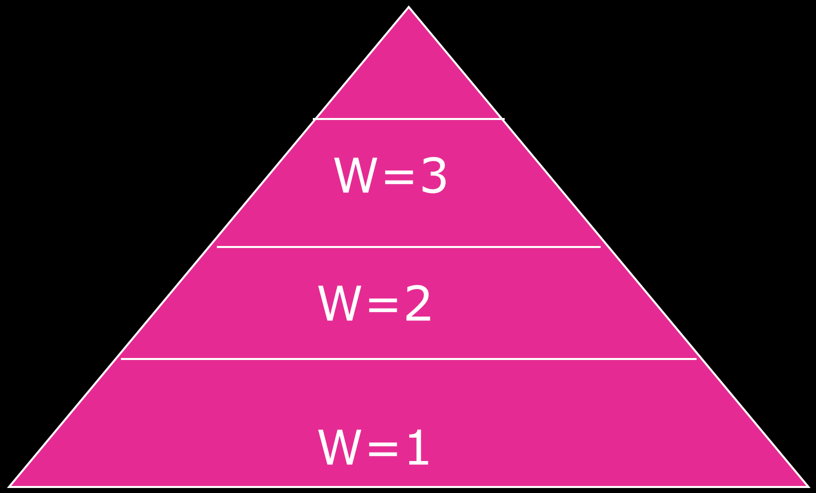
A. The cause of the fire was unknown.

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

Running a pyramid evaluation on mechanical turk

- Use crowd sourcing to build pyramids
  - Ask turkers to select 8 simple sentences from human summaries - > 13 SCUs/summary
- Lightweight sampling for evaluating system output
  - From 51 SCUs/reference summaries, sample 32
    - Why would sampling work?
  - Ask turkers if a pyramid SCU appears in system generated summary

Shapira et al, NAACL 2019

Automating the pyramid method

- Decomposes summaries into clauses using parsing
- Represents each clause as a semantic vector
- Uses a set partitioning algorithm to maximize similarity between clauses and group into SCUs.
- Semantic similarity computed between system output and pyramids

Gao et al, EMNLP 2019

Next Time

- Neural Net approaches to summarization