Summarization

Summarization

- *Text summarization is the process of distilling the most important information from a text to produce an abridged version for a particular task and user*
  - Definition adapted from Mani and Maybury 1999
- Types of summaries in current research:
  - Outlines or abstracts of any document, article, etc.
  - Snippets summarizing a Web page or a search engine results page
  - Action items or other summaries of a business meeting
  - Summaries of email threads
  - Simplifying text by compressing sentences
  - Keyword extraction

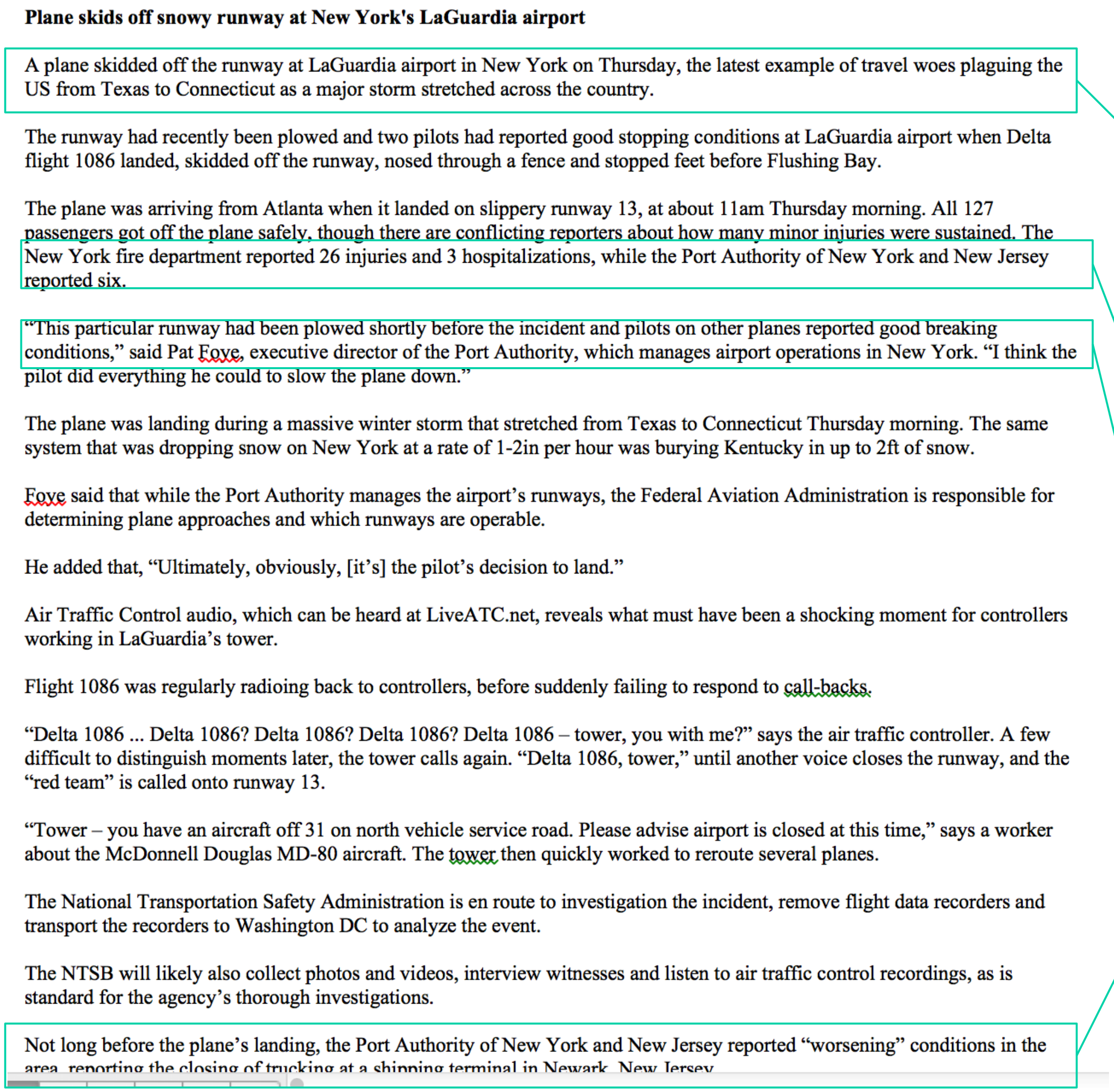
Single vs. Multiple Documents

- Single-document summarization
  - Given a single document, produce a gist of the content in the form of an abstract or outline
- Multiple-document summarization
  - Given a group of documents, produce a gist of the content, and create a cohesive answer that combines information from each document
    - a series of news stories on the same event
    - a set of web pages about some topic or question

Extractive vs. Abstractive

- Abstractive summarization:
  - express the ideas in the source documents using (at least in part) different words
  - how humans typically approach summarization
- Extractive summarization:
  - create the summary from phrases or sentences in the source document(s)

Example of Extractive Summary



Plane skids off snowy runway at New York's LaGuardia airport

A plane skidded off the runway at LaGuardia airport in New York on Thursday, the latest example of travel woes plaguing the US from Texas to Connecticut as a major storm stretched across the country.

The runway had recently been plowed and two pilots had reported good stopping conditions at LaGuardia airport when Delta flight 1086 landed, skidded off the runway, nosed through a fence and stopped feet before Flushing Bay.

The plane was arriving from Atlanta when it landed on slippery runway 13, at about 11am Thursday morning. All 127 passengers got off the plane safely, though there are conflicting reporters about how many minor injuries were sustained. The New York fire department reported 26 injuries and 3 hospitalizations, while the Port Authority of New York and New Jersey reported six.

"This particular runway had been plowed shortly before the incident and pilots on other planes reported good breaking conditions," said Pat Foye, executive director of the Port Authority, which manages airport operations in New York. "I think the pilot did everything he could to slow the plane down."

The plane was landing during a massive winter storm that stretched from Texas to Connecticut Thursday morning. The same system that was dropping snow on New York at a rate of 1-2in per hour was burying Kentucky in up to 2ft of snow.

Foye said that while the Port Authority manages the airport's runways, the Federal Aviation Administration is responsible for determining plane approaches and which runways are operable.

He added that, "Ultimately, obviously, [it's] the pilot's decision to land."

Air Traffic Control audio, which can be heard at LiveATC.net, reveals what must have been a shocking moment for controllers working in LaGuardia's tower.

Flight 1086 was regularly radioing back to controllers, before suddenly failing to respond to call-backs.

"Delta 1086 ... Delta 1086? Delta 1086? Delta 1086? Delta 1086 – tower, you with me?" says the air traffic controller. A few difficult to distinguish moments later, the tower calls again. "Delta 1086, tower," until another voice closes the runway, and the "red team" is called onto runway 13.

"Tower – you have an aircraft off 31 on north vehicle service road. Please advise airport is closed at this time," says a worker about the McDonnell Douglas MD-80 aircraft. The tower then quickly worked to reroute several planes.

The National Transportation Safety Administration is en route to investigation the incident, remove flight data recorders and transport the recorders to Washington DC to analyze the event.

The NTSB will likely also collect photos and videos, interview witnesses and listen to air traffic control recordings, as is standard for the agency's thorough investigations.

Not long before the plane's landing, the Port Authority of New York and New Jersey reported "worsening" conditions in the area, reporting the closing of trucking at a shipping terminal in Newark, New Jersey

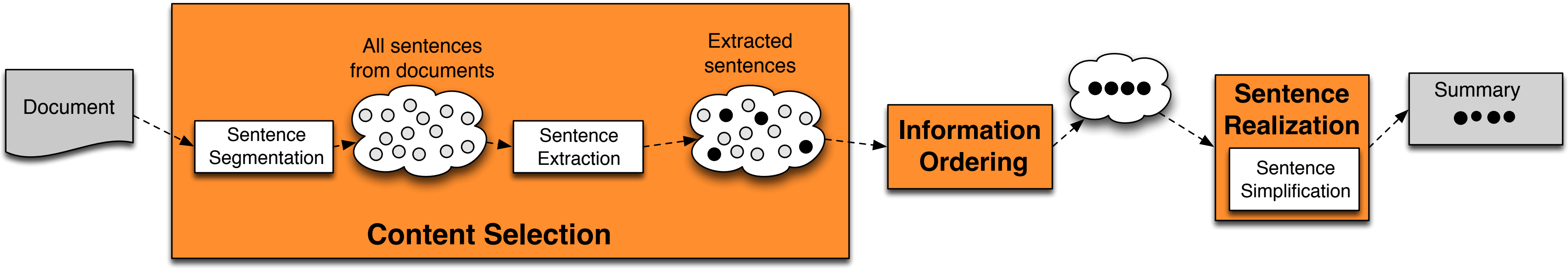
Plane skids off snowy runway at New York's LaGuardia airport

A plane skidded off the runway at LaGuardia airport in New York on Thursday, the latest example of travel woes plaguing the US from Texas to Connecticut as a major storm stretched across the country. The New York fire department reported 26 injuries and 3 hospitalizations, while the Port Authority of New York and New Jersey reported six.

"This particular runway had been plowed shortly before the incident and pilots on other planes reported good breaking conditions," said Pat Foye, executive director of the Port Authority, which manages airport operations in New York. Not long before the plane's landing, the Port Authority of New York and New Jersey reported "worsening" conditions in the area, reporting the closing of trucking at a shipping terminal in Newark, New Jersey.

Typical approaches to general problem

- Currently, achieve extraction instead of a true re-phrasing
  - **Content Selection**
    - Identify the sentences or clauses to extract
  - **Information Ordering**
    - How to order the selected units
  - **Sentence Realization**
    - Perform cleanup on the extracted units so that they are fluent in their new context
      - E.g. replacing pronoun or other references left dangling



Content Selection: Centrality Methods

- Centrality methods select sentences based on properties of words and sentences
  - Doesn’t require manual training data
- Simple approach is to select sentences that have more informative words according to saliency defined from a topic signature of the document
- Centroid-based summarization uses log-likelihood ratios for words, computing the probability of observing the word in the input more often than in the background corpus
- TextRank or LexRank method scores the importance of sentences in a similar way to PageRank on web pages
  - Each sentence is a vertex of a graph and the PageRank algorithm assigns scores based on the similarity of the words in the sentences

Content Selection: Other Methods

- Methods based on rhetorical parsing use coherence relations to identify satellite and nucleus sentences
- Machine learning methods score sentences for importance from a corpus of sentences assigned importance scores
  - use features based on
    - Position,
    - cue phrases,
    - word informativeness,
    - sentence length,
    - cohesion (computing lexical chains of the document)

Information Ordering

- Single documents can simply keep the document ordering
- Chronological ordering:
  - Order sentences by the date of the document (for summarizing news)..

  (Barzilay, Elhadad, and McKeown 2002)
- Coherence:
  - Choose orderings that make neighboring sentences similar (by cosine).
  - Choose orderings in which neighboring sentences discuss the same entity (Barzilay and Lapata 2007)
- Topical ordering
  - Learn the ordering of topics in the source documents

Simplifying Sentences

Zajic et al. (2007), Conroy et al. (2006), Vanderwende et al. (2007)

- Simplest method: parse sentences, use rules to decide which modifiers to prune
  - (more recently a wide variety of machine-learning methods)

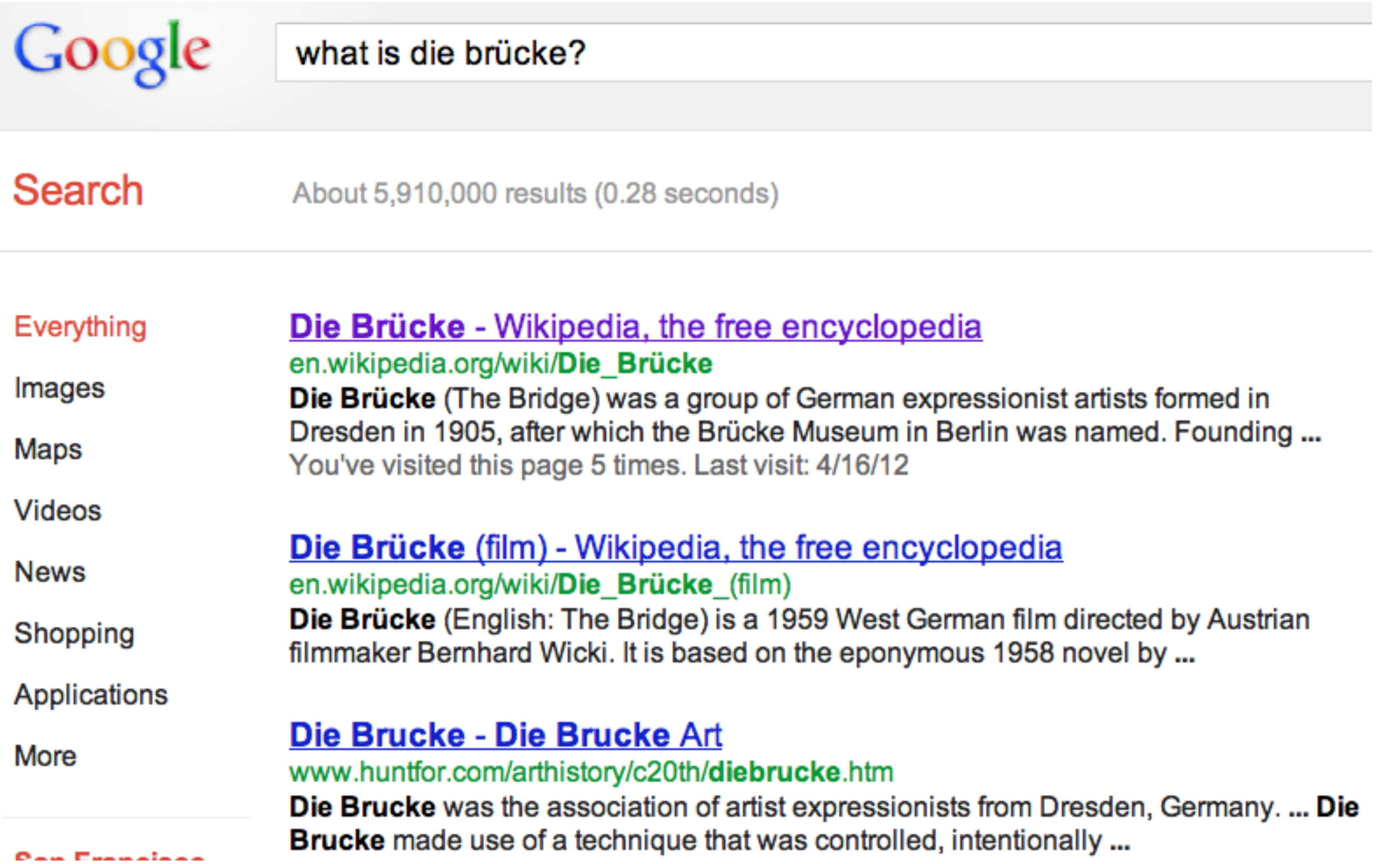
appositives	Rajam, ~~28, an artist who was living at the time in Philadelphia~~, found the inspiration in the back of city magazines.
attribution clauses	Rebels agreed to talks with government officials, ~~international observers said Tuesday.~~
PPs **without named entities**	The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [~~PP to a sustainable number~~]]
initial adverbials	"~~For example~~", "~~On the other hand~~", "~~As a matter of fact~~", "~~At this point~~"

Summarization Evaluation

- Extrinsic (task-based) evaluation: humans are asked to rate the summaries according to how well they are enabled to perform a specific task
- Intrinsic (task-independent) evaluation
  - Human judgments to rate the summaries
  - ROUGE (Recall Oriented Understudy Gisting Evaluation)
    - Humans generate summaries for a document collection
    - System-generated summaries are rated according to how close they come to the human-generated summary
    - Measures have included unigram overlap, bigram overlap, and longest common subsequence
  - Pyramid method
    - Humans identify “units of meaning” and then an overlap measure is computed

Summarization for Web Search: Snippets

- **Create** **snippets** summarizing a web page for a query
  - Google: 156 characters (about 26 words) plus title and link



what is die brücke?

About 5,910,000 results (0.28 seconds)

Die Brücke - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Die_Brücke**
Die Brücke (The Bridge) was a group of German expressionist artists formed in Dresden in 1905, after which the Brücke Museum in Berlin was named. Founding ...
You've visited this page 5 times. Last visit: 4/16/12

Die Brücke (film) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Die_Brücke**_(film)
Die Brücke (English: The Bridge) is a 1959 West German film directed by Austrian filmmaker Bernhard Wicki. It is based on the eponymous 1958 novel by ...

Die Brucke - **Die Brucke** Art
www.huntfor.com/arthistory/c20th/**diebrucke**.htm
Die Brucke was the association of artist expressionists from Dresden, Germany. ... **Die Brucke** made use of a technique that was controlled, intentionally ...