Natural Language Processing
SoSe 2014

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Summarization

Dr. Mariana Neves

June 25th, 2014

(based on the book of Jurafski and Martin 2009)

Outline

- Task
- Single-document summarization
- Multi-document summarization
- Query-focused summarization
- Evaluation

Outline

- Task
- Single-document summarization
- Multi-document summarization
- Query-focused summarization
- Evaluation

Summarization

- Half-way between
  - Information retrieval (entire documents)
  - Question answering (factoid answers)

- „It is the process of distilling the most important information from a text to produce an abridged version for a particular task and user“ (Jurafski and Martin 2009)

Summarization

- Kinds of summaries
  - Outlines of a document
  - Abstracts of a scientific article
  - Headlines of news articles
  - Snippets summarizing a Web page on a search engine results page
  - Action items or other summaries of a (spoken) business meeting
  - Summaries of emails threads
  - Answers to complex questions (multi-documents)

Summarization

- Dimensions
  - Single-document
    - Headlines of new articles, abstracts of scientific publications
  - Multiple-document
    - Series of new stories of the same event, emails from a thread
  - Generic: focus of the important information of the document(s)
  - Query-focused
    - Question answering

Abstract vs. Extract

- Extract
  - Combination of phrases and sentences from the document(s)
- Abstract
  - Uses different words to describe the content of the document(s)
- Most current summarizers are extractive (easier)

Abstract vs. Extract

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation or any nation so conceived and so dedicated can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead who struggled here have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us the living rather to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us — that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that we here highly resolve that these dead shall not have died in vain, that this nation under God shall have a new birth of freedom, and that government of the people, by the people, for the people shall not perish from the earth.

Figure 1.1: The Gettysburg Address

(figure taken from Mani 2001)

Abstract vs. Extract

> Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. **The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract.**

Figure 1.3: Another 25% extract of the Gettysburg Address

> **This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.**

Figure 1.4: A 15% abstract of the Gettysburg Address

(figures taken from Mani 2001)

Architecture of summarization systems

- Content selection
  - Usually sentences and clauses
- Information ordering
  - Order and structure the extracted units
- Sentence realization
  - Clean up to assure fluency

Outline

- Task
- Single-document summarization
- Multi-document summarization
- Query-focused summarization
- Evaluation

Single-document summarization

- Content selection
  - Choose sentences
  - Binary classification task
    - Important (extract worthy)
    - Unimportant (not extract worthy)

- Information ordering
  - Sentences are ordered by their original order in the document

- Sentence realization
  - Remove non-essential phrases from the sentences
  - Fusing sentences into a single one

Unsupervised content selection

- Select sentences with more salient or informative words
- Saliency
  - Topic signature: set of salient or signature terms with salient scores greater than a threshold θ
- Weight schemes instead of word frequencies
  - Tf-idf

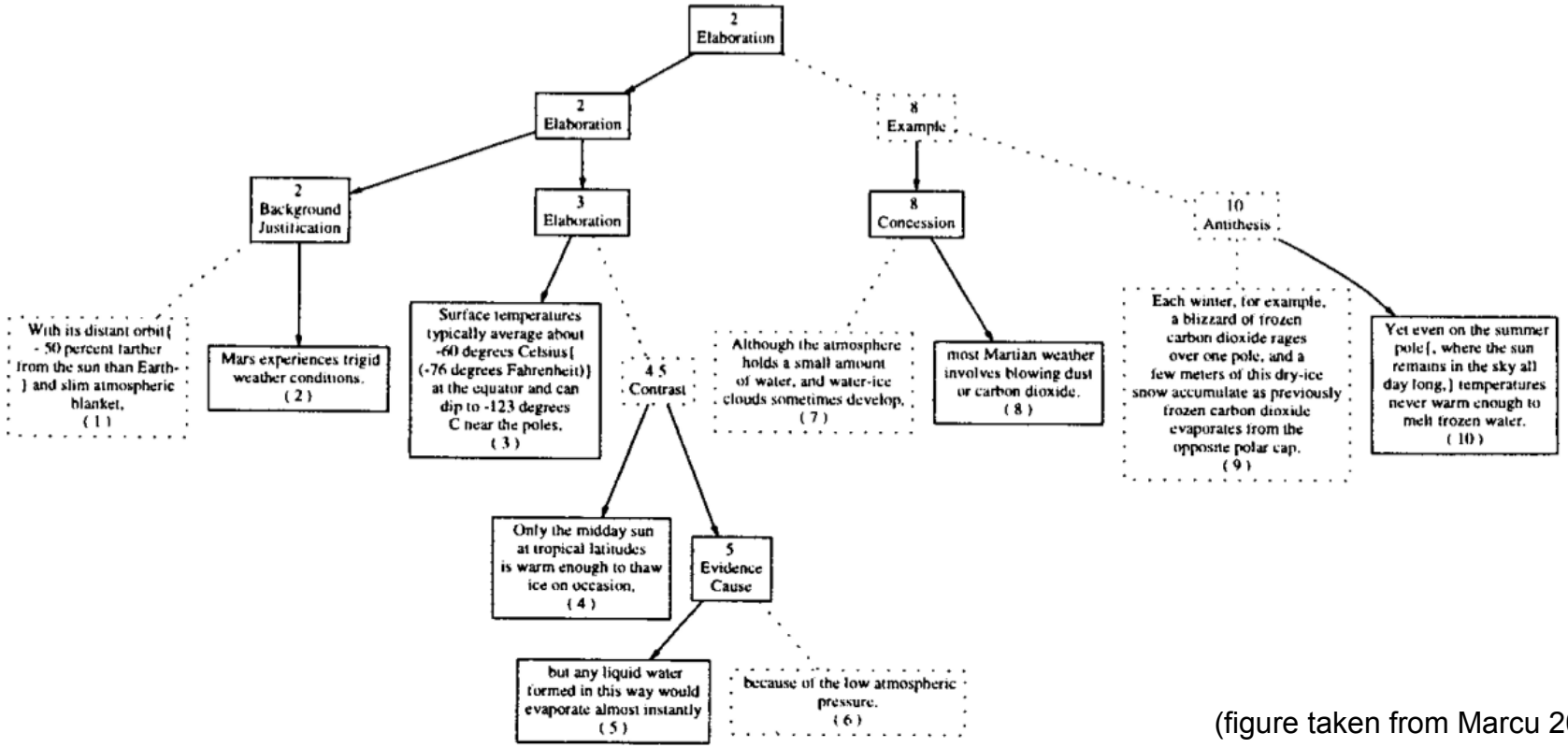
Centroid-based summarization

- Set of signature terms as a pseudo-sentence that is the „centroid" of all sentence in the document
- We look for sentences which are close to this centroid sentence
- Compute distances between each candidate sentence *x* and each other sentence *y*
- Choose sentences which are on average closer to other sentences

$$centrality(x) = \frac{1}{K} \sum_{y} tf.idf.cosine(x, y)$$

Rhetorical parsing

- Introduce more sofisticated discourse knowledge
- Applying a discourse parser to compute coherence relations between the discurse units



(figure taken from Marcu 2000)

Supervised Content Selection

- Efectivelly combine various features from the text
- Training data: documents and respective summaries
- Extracts of sentences:
  - Classification task: 1 (present); 0 (not present)

Supervised Content Selection

- Features
  - Position of the sentence in the text:
    - Title
    - First sentence of paragraph 2
    - First sentence of paragraph 3
    - Final sentence
  - Cue phrases
    - „In summary..“, „In conclusion..“, „This paper..“
  - Word informativeness
    - Topic signature

Supervised Content Selection

- Features
  - Sentence length
    - Long sentences rather than short ones
    - Binary feature based on a cutoff (e.g., 5 words)
  - Cohesion
    - Sentences that contain more terms from a lexical chain (series of related words) are extract worthy
    - Can also be computed using graph-based methods (e.g., PageRank)

Supervised Content Selection

- Using abstracts of documents as training data
- Need to align sentences in abstracts to the document text
  - Longest common subsequences of non-stopwords
  - Minimum edit distance

Sentence realization

- Sentence compression or sentence simplification
- Running a syntactic parser and prunning some phrases
- Examples:
  - Apposition: „Barry Goldwater, ~~the junior senator from Arizona~~, received the Republican nomination in 1964“
  - Attribution clauses: „Rebels agreed to talks with governments, ~~international observers said Tuesday~~ “
  - PPs w/o NERs
  - Initial adverbials: „For example“, „On the other hand“, „At this point“, etc.
- Also supervised machine learning

Outline

- Task
- Single-document summarization
- Multi-document summarization
- Query-focused summarization
- Evaluation

Multi-document summarization

- Applications
  - Summarize Web pages for a particular event in the news
  - Finding answers to complex questions
- Architecture
  - Content selection
  - Information ordering
  - Sentence realization
- Use of supervised methods over supervised ones
  - Not much training data available

Content selection (Multi-doc)

- Redundancy of information
- Summaries should not be consisted of identical or similar sentences
- Calculating the redundancy factor between new extracted sentences and current selected sentences
- Maximal Marginal Relevance (MMR)
  - λ is a weight that can be tuned
  - Similarity is some similarity function

$$MMR\ penalization\ factor(s) = \lambda\ max_{s_i \in Summary}\ Similarity(s, s_i)$$

Content selection (Multi-doc)

- Clustering algorithm
  - Groups sentences in clusters of related sentences
  - Select a single (centroid) sentence from each cluster

- Sentence simplification or compression in this step
  - Produde many variations of the original sentence
  - Let the clustering or MMR select the best one

Information ordering (Multi-doc)

- Concatenate extracted sentences in a coherent way

- Chronological ordering
  - If date of the original document/article is available (e.g, news)
  - But usually lack cohesion

- Coherence
  - Coherence relations between sentences
  - Cohesion and lexical chains (local cohesion)

Information ordering (Multi-doc)

- Lexical cohesion
  - Ordering sentences next to sentences which contain similar words
  - tf.idf, cosine similarity between pair of sentences
  - Minimizing distance between neighboring sentences
- Centering
  - Salient entities
  - Syntactic realization of the focus (i.e., subject or object)
  - Transitions between realizations

Information ordering (Multi-doc)

1 [The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$.
2 [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$.
3 [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.
4 [Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically.
5 [The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation of the Sherman Act]$_O$.
6 [Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

(figure taken from Barzilay and Lapata 2005)

Information ordering (Multi-doc)

- Given coherence score for pairs or sequence of sentences
- Problem: find the optimal ordering of sentences
- NP-complete
  - But there are good approximation methods
  - Althaus et al. 2004, Knight 1999, Cohen et al 1999, Brew 1992

Sentence realization (Multi-doc)

- Checking further for coherence
- Longer or more descriptive phrases should come before short, reduced or abbreviated forms
- Examples
  - „U.S. President George W. Bush“ and „Bush“
- Co-reference resolution algorithm
- Rewrite, cleanup rules

Sentence realization (Multi-doc)

Original summary:
Presidential advisers do not blame **O'Neill**, but they've long recognized that a shakeup of the economic team would help indicate **Bush** was doing everything he could to improve matters. **U.S. President George W. Bush** pushed out **Treasury Secretary Paul O'Neill** and top economic adviser Lawrence Lindsey on Friday, launching the first shake - up of his administration to tackle the ailing economy before the 2004 election campaign.

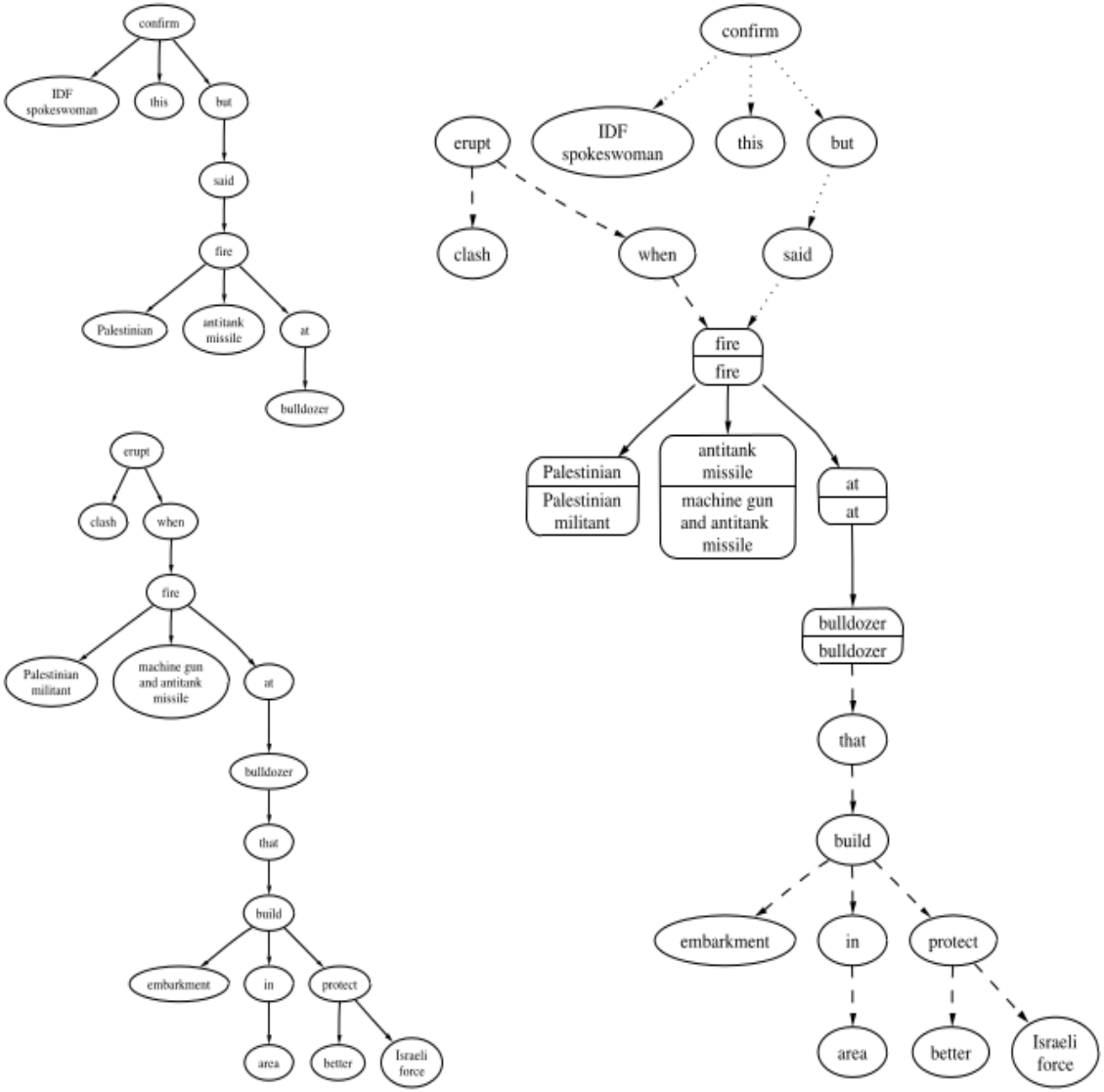
Rewritten summary:
Presidential advisers do not blame **Threasury Secretary Paul O'Neill**, but they've long recognized that a shakeup of the economic team would help indicate **U.S. President George W. Bush** was doing everything he could to improve matters. **Bush** pushed out **O'Neill** and White House economic adviser Lawrence Lindsey on Friday, launching the first shake-up of his administration to tackle the ailing economy before the 2004 election campaign.

(figure taken from
Nenkova and McKeown 2003)

Sentence realization (Multi-doc)

- Sentence fusion
  - Parsing each sentence
  - Alignment of the parses to find common information
  - Build a fusion structure with overlapping information
  - Create a new fused sentence

Sentence realization (Multi-doc)



(figure taken from Barzilay and McKeown 2005)

Outline

- Task
- Single-document summarization
- Multi-document summarization
- Query-focused summarization
- Evaluation

Query-focused summarization

- Question answering
  - Longer, descriptive, more informative answers
- Example: (BioASQ training data)
  - "What is the function of the mammalian gene Irg1?"
  - "Human IRG1 and mouse Irg1 mediates antiviral and antimicrobial immune responses, without its exact role having been elucidated. Irg1 has been suggested to have a role in apoptosis and to play a significant role in embryonic implantation. Irg1 is reported as the mammalian ortholog of methylcitrate dehydratase."

Query-focused summarization

- Content selection
  - Adapt multi-doc content selection to rank sentences based relevance to the query
    - Overlapping words query/sentences
    - Cosine similarity query/sentence
  - Build a top-down expectations for each topic
    - Biography: dates, nationalities, educations, etc.
    - Drug efficacy: population, problem/disease, intervention, outcome, side-effects, etc.

Query-focused summarization

- Content selection
  - Use of templates:
    - Example: Biography

      <NAME> is <WHY_FAMOUS>.

      She/He was born on <BIRTH_DATE> in <BIRTH_LOCATION>.

      She/He <EDUCATION>.

      <DESCRIPTIVE_SENTENCE>

      <DESCRIPTIVE_SENTENCE>

      ...

Outline

- Task
- Single-document summarization
- Multi-document sumamrization
- Query-focused summarization
- Evaluation

Evaluation

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  - Measure the amount of overlapping N-grams between automatic and human-generated summaries
  - ROUGE-1 (unigram), ROUGE-2 (bigram), etc.

$$ROUGE2 = \frac{\sum_{S \in Summaries} \sum_{bigrams \in S} Count_{match}(bigram)}{\sum_{S \in Summaries} \sum_{bigrams \in S} Count(bigram)}$$

Evaluation

- ROUGE
  - Recall-oriented measure
  - ROUGE-L
    - Longest common subsequence
  - ROUGE-S, ROUGE-SU
    - Skip bigrams: pair of words in a certain order by allowing any number of words between them

Further Reading

- Speech and Language Processing
  - Chapters 23.3 – 23.8