

COMP348 — Document Processing and the Semantic Web

Week 10 Lecture 1: Document Summarisation

Diego Mollá

COMP348 2008H1

Abstract

We will address the issues related with summarisation and will focus on one kind of summarisation, extractive summarisation.

Update May 14, 2008

Contents

1 Document Summarisation	1
2 Document Summarisation Techniques	3
2.1 Sentence Extraction	3
2.2 Cohesion Check	5
2.3 Balance and Coverage	7

Some Useful Extra Reading

- E. H. Hovy, Automated Text Summarization, In: R. Mitkov (ed), *The Oxford Handbook of Computational Linguistics*, pp. 583–598, 2003. <http://online.mq.edu.au/pub/COMP348/reading/hovy.pdf>

1 Document Summarisation

What is Document Summarisation

Summarisation (or automatic abstracting)

A summary is a text that is produced from one or more texts, that contains a significant portion of the information of the original text(s), and that is no longer than half of the original text(s). (Hovy, 2003)

What is Document Summarisation Good For?

What for?

- For busy people to read the summary instead of the full text → *informative summary*
- For researchers, web surfers, . . . to read the summary to decide if it is worth to read the original text → *indicative summary*
- To avoid having to type out an abstract for a technical report when the publisher requests it → not realistic

Examples

Original Text

Abraham Lincoln was born in Kentucky on February 12th, 1809. His family moved to Indiana when he was eight years old. His mother died when he was ten. Abraham Lincoln loved to read. He only went to school for a few months. He had to study and learn on his own. From his childhood, Abraham Lincoln was known as a hard worker. He worked on a farm and in a store and on a boat. He studied law and practiced for many years in Illinois. He married Mary Todd Lincoln. Together they had four sons. Abraham Lincoln was elected as the 16th President of the United States in 1860. He did many things as President. Many people think he was the best American President of all time. He is most remembered for freeing the slaves. He was President of the United States during the time the Civil War was fought. The Civil War was fought between the Northern and Southern states. He was known as a great speaker. Some of his most famous speeches include the Emancipation Proclamation - in which he freed the slaves - and the Gettysburg Address, which starts "Four score and seven years ago ..." While attending a play on April 14th, 1865, Abraham Lincoln was shot by John Wilkes Booth. Lincoln died the next morning. Lincoln's birthday is celebrated in February of each year.

Word's Autosummarize

Abraham Lincoln was born in Kentucky on February 12th, 1809.

Abraham Lincoln loved to read.

Abraham Lincoln was elected as the 16th President of the United States in 1860.

The Civil War was fought between the Northern and Southern states.

While attending a play on April 14th, 1865, Abraham Lincoln was shot by John Wilkes Booth.

Lincoln died the next morning.

MEAD

Abraham Lincoln was born in Kentucky on February 12th 1809.

His family moved to Indiana when he was eight years old.

From his childhood Abraham Lincoln was known as a hard worker.

Abraham Lincoln was elected as the 16th President of the United States in 1860.

While attending a play on April 14th 1865 Abraham Lincoln was shot by John Wilkes Booth.

Summarisers on the Web

- Integrate a summariser into your Web browser:
 - <http://www.copernic.com/en/products/summarizer/>
 - <http://www.kryltech.com/summarizer.htm>
- Information on summarisation on the Web:
 - <http://www.summarization.com/>
- Web Demo:
 - MEAD <http://tangra.si.umich.edu/clair/md/demo.cgi>

2 Document Summarisation Techniques

An Ideal Document Summarisation System

Understanding Stage

Document(s) → Knowledge base

Generation Stage

Knowledge base → Summary

A Compromise Solution

Sentence Extraction

Document → Sentence candidates

This is what most commercial and free summarisers do

Cohesion Check

Sentence candidates → Coherent text

Balance and Coverage

Coherent text → Summary

2.1 Sentence Extraction

General Approach

For each sentence

1. Look for clues to its importance
2. Compute a score for the sentence based on the clues found
3. Select all sentences whose score exceeds some threshold
 - Or select the highest scoring sentences up to a certain total

The Frequency-keyword Approach

1. Compute the *keywords* of the document:
 - ignore the function words by using a stop word list
 - sort all remaining words according to frequency
 - select the most frequent words (say, the top 5%).
2. Score the document sentences according to the presence of keywords:
 - simple keyword count
 - weighted keyword count (keyword weights for each sentence)
 - looking for keyword clusters in the sentence

The Biased Keyword Approach

Title and headings biased

Compute a list of keywords on the basis of document structure:

- select candidates from titles and headings only, or
- candidates from titles and headings have more importance:
 - e.g. they are counted as being more frequent.

Query biased (customised summaries)

Use the user's query to determine the keyword's weights:

- the user's query determines all the keywords, or
- the user's query introduces additional keywords or updates the weights of existing keywords.

The Location Method

Observation

First and last sentence of a paragraph are usually most central to the theme of a text

Increase the score of a sentence according to its position in the paragraph:

- beginning of paragraph
- end of paragraph

Syntactic Criteria

Method

- Use the syntactic structure to determine the sentence's importance
- Tag the words according to their part of speech
- Alternatively, use a parser to find the sentence structure

Problem

- there are too many types of syntactic structures to consider
- often a very high percentage of the patterns occurs only once in the corpus (sparse data)

Cues, Indicator Phrases

Cues

- Certain words (not necessarily keywords) provide an indication of the importance of the sentence
- Use these words to determine the sentence score:
 - *bonus words* increase the sentence score:
 - * “greatest”, “significant”
 - *stigma words* decrease the sentence score:

- * “hardly”, “impossible”, “now”

Indicator Phrases

Indicator phrases are specific phrases or patterns of phrases that can be used to determine the sentence importance:

- “The main aim of the present paper is ...”
- “The purpose of this article is ...”
- “In this report, we outline ...”
- “Our investigation has shown that ...”

Relational Criteria

1. Build a semantic structure for the document:
 - sentences are vertices
 - inter-sentence links are edges
 - Rhetorical links (elaboration, sequence, etc)
 - Cooccurrence of keywords
 - ...
2. Use a centroid metric to determine the most important sentences
 - Degree of the vertex
 - Eigenvalues (PageRank style)

2.2 Cohesion Check

Textual Cohesion

- Lack of cohesion results in “odd” extracts
- Sentences include references to other sentences:
 - Anaphoric reference:
 - “John saw Mary. *She* was talking over the phone”
 - Rhetorical connectives:
 - “*So*, the following example ...”
 - Lexical or definite reference:
 - “I saw a man with a book. *The book* was ...”

- Possible solutions:

Aggregation Add preceding sentences until there are no external references

Deletion Remove the difficult sentences

Modification Alter the sentences to eliminate or disguise the problem

Anaphoric Reference

- Anaphors are the words or phrases that refer to other parts of the text

Internal Anaphora The referent is in the same sentence

External Anaphora The referent is in another sentence

- Potentially anaphoric words are not always used anaphorically
- An anaphor recognition algorithm is necessary
 - Detect anaphoric words
 - Find the referent

Definite Noun Phrases

- Noun phrases introduced by the definite article (“the”) are regarded as typically referential:
 - “the book” refers to a book that has been introduced before
- There are non-referential uses of definite noun phrases:
 - “the” plus one or two words plus “of”, “between”, “in”
 - They can describe permanently available referents
 - * “the sun”, “the weather”, “the future”, “the human race”
 - * “the literature”, “the periodic table” (domain-dependent)
 - They can describe well-known organisations or persons
 - * “the United States”, “the President” (clue: usually capitalised)

Logical and Rhetorical Connectives

- These connectives indicate the nature of the relationship between a sentence and its predecessor or successor
- Some of them are unambiguous:
 - “but”
 - “in fact”
 - “nonetheless”
 - “however”
 - “on the other hand”
 - “indeed”

Aggregation

We need to tidy-up non-coherent sentences

Backward Aggregation

- Add previous sentences one by one until there are no external references
- Rationale: references are almost always backwards
- Definite references may have long-distance referents
- We may have to add several sentences in one go

Forward Aggregation

- Several rhetorical connectives in the next sentence indicate further elaboration (“in fact”, “in particular”, “thus”)
- Add the next sentence to make the abstract more readable

2.3 Balance and Coverage

Balance and Coverage

- We need to process the selected sentences in order to produce a real abstract:
 - Delete redundant sentences
 - Harmonise tense and voice of verbs
 - Ensure balance and proper coverage
- Combination of information extraction and text generation
- Need to consider text structure:
 - Each sentence plays a role in the text and in relation with the other sentences
- Problem to address:
 - Lack of balance and coverage:
 - * Missing important information
 - * Too much emphasis on less important information

Textual Structure

- Two ways of producing abstracts:
 - As freely-formatted text
 - As formatted abstracts
- The structure of an abstract reflects the structure of the text
- Freely-formatted abstracts do not make this structure explicit
- Formatted abstracts make this structure explicit:
 - “Purpose of study”, “Procedures”, “Findings”, etc.

Abstract-frames

- Abstract-frames provide a template for the abstract
- Fill the blanks with information extracted from the document
- This is a variant of information extraction
- A small number of abstract-frames can be used for a wide range of (technical) literature
- Try to fill at least the most important slots
- Tailored abstracts: the user may influence the determination of the importance of each slot

Slot Instantiation

Procedure

1. Classify each portion of the text according to the slot it belongs:
 - Look at headings
 - Look for cues and indicator phrases
 - Use statistical (machine learning) methods
2. For each slot, select the most representative text:
 - According to the importance of each piece of text with respect to the slot

Examples of Cues and Indicator Phrases

<i>Cue, Indicator Phrase</i>	<i>Classification</i>
“We have shown that . . .” past tense forms and citations “made”, “used”, “measured”, “determined”	findings historical background experimental methods

Coverage

- Abstract frames:
 - They ensure better balance, but . . .
 - They do not ensure better coverage
- How to ensure better coverage:
 - Customise the abstract frames to match the document structure
 - Separate sections deal with separate aspects of a document’s message

Take-home Messages

Essential Skills — for a passing grade

- Explain the three main stages of summarisation
- Explain a significant number of the main approaches to sentence extraction

Additional Skills — for credit or above

- Explain in detail all main approaches to sentence extraction
- Explain the role of anaphors in textual cohesion and how to handle them
- Explain the use of abstract frames to ensure good balance

What's Next

Week 10

- Lecture 2: Information Extraction
- Lecture 3: Named Entity Recognition