CSE528

Natural Language Processing

Venue:ADB-405 SLOTS: A2+TA2 Topic: Text Summarization

Prof. Tulasi Prasad Sariki,

SCSE, VIT Chennai Campus

www.learnersdesk.weebly.com

Contents

- What is Text Summarization
-

Rapid growth of data

The problem:

- 4 Billion URLs indexed by Google
- 200 TB of data on the Web [Lyman and Varian 03]

Possible approaches:

- information retrieval
- document clustering
- information extraction
- visualization
- question answering
- text summarization

Text Summarization

Automatic Text Summarization

- No User interaction, system will return the condensed or summarized form.

Query Specific Summarization

- User interaction will be there he/she will supply some input as keywords based on that summary will be generated.

Automatic Text Summarization

Automatic text summarization is the technique where a computer automatically creates an abstract or gist of one or more text documents.

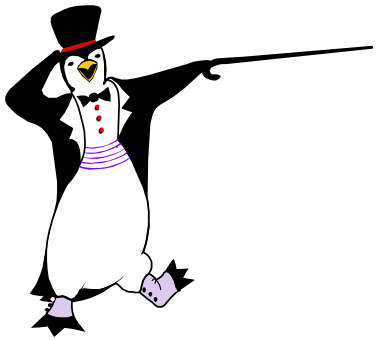
Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users)and task (or tasks).

Typical 3 Stages of Summarization

1. Topic Identification: find/extract the most important material

2. Topic Interpretation: compress it

3. Summary Generation: say it in your own words



…as easy as that!

Aspects that describe Summaries

Source (Input):

- Source: single-document vs. multi-document
- Language: mono-lingual vs. multi-lingual vs. cross-lingual
- Genre: news vs. technical report vs. scientific paper etc.
- Specificity: domain-specific vs. general
- Length: short (1–2 pages) vs. long (> 2 pages)
- Media: text, graphics, audio, video, multi-media etc.

Aspects that describe Summaries

Purpose:

- Use: generic vs. query-oriented (aimed to a specific information need)
- Purpose: what the summary is used for (e.g. alert, preview, inform, digest, provide biographical information)
- Audience: untargeted vs. targeted (aimed at a specific audience)

Composition (Output):

- Derivation: extract vs. abstract
- Format: running text, tables, geographical displays, time lines, charts, illustrations etc.
- Partiality: neutral vs. evaluative (adding sentiment/values)

Query-Driven vs. Text-Driven

Top-down: Query-driven focus

- *Criteria of interest* encoded as search specs.
- System uses specs to filter or analyze text portions.
- Examples: *templates* with slots with semantic characteristics; *term lists* of important terms.

Bottom-up: Text-driven focus

- *Generic importance metrics* encoded as strategies.
- System applies strategies over rep of whole text.
- Examples: degree of *connectedness* in semantic graphs; *frequency* of occurrence of tokens.

Extract not Abstract

Extraction is much easier than abstraction

Abstraction needs understanding and rewriting

Most automatic summarization tools makes extracts not abstracts

Uses original sentences or part of sentences to create ”abstract”

Some Extraction Methods

General method: score each sentence; choose best sentence(s)

Scoring techniques:

- Position in the text: lead method; optimal position policy; title/heading method
- Cue phrases in sentences
- Word frequencies throughout the text
- Cohesion: links among words; word co-occurrence; coreference; lexical chains
- Information Extraction: parsing and analysis

Word Frequency[Luhn58]

Steps:

- Count all word occurrences (after stemming)
- Ignore extreme frequencies.
- Give every word a score according to frequency.
- Calculate the importance of each sentence as the sum of its word scores.
  - Take the physical distance between important words into consideration.
- Extract the N sentences with the highest scores.

Position: Title-Based Method

Words in titles and headings are positively relevant to summarization.

Shown to be statistically valid at 99% level of significance (Edmundson, 68).

Empirically shown to be useful in summarization systems.

Cue words and phrases

Baxendale (1958) identified two sets of phrases

- bonus phrases -> that tend to signal when a sentence is a likely candidate for inclusion in a summary
- stigma phrases -> that tend to signal when it is definitely not a candidate, respectively.

‘Bonus phrases’ such as "in summary", "in conclusion", and superlatives such as "the best", "the most important" can be good indicators of important content.

‘stigma phrases’ such as *hardly* and *impossible* may indicate non-important sentences

Cue words and phrases

Cue words and phrases, such as "in conclusion", "important", "in this paper", "this paper", "this article", "this document", and "we conclude“ etc. can be very useful to determine signals of relevance or irrelevance.

During processing, the Cue Phrase Module simply rewards each sentence containing a cue phrase with an appropriate score (constant per cue phrase) and penalizes those containing stigma phrases.

Multiple Methods

Cue-Phrase Method: Some phrases imply significance: “significant”, “impossible”, ”hardly”, etc.

Key Method: Word frequencies, like Luhn(for ATS).

Title Method: Titles are important, and so are the words they contain sentences are play major role in summary.

Location Method: First and Last sentences of a paragraph, sentences following titles.

Multiple Methods

The Sentence importance is calculate as a linear combination of the different methods:

Sentence Score = ß1 Cue + ß2 Key + ß3 Title + ß4 Location.

Adjust the coefficients to control each methods significance.

Cohesion: Lexical chains method

But Mr. Kenny's move speeded up work on a **machine** which uses **micro-computers** to control the rate at which an *anaesthetic* is pumped into the blood of *patients* undergoing *surgery*. Such **machines** are nothing new. But Mr. Kenny's **device** uses two **personal-computers** to achieve much closer monitoring of the **pump** feeding the *anaesthetic* into the *patient*. Extensive testing of the **equipment** has sufficiently impressed the authorities which regulate *medical* **equipment** in Britain, and, so far, four other countries, to make this the first such **machine** to be licensed for commercial sale to *hospitals*.

Lexical chains-based method

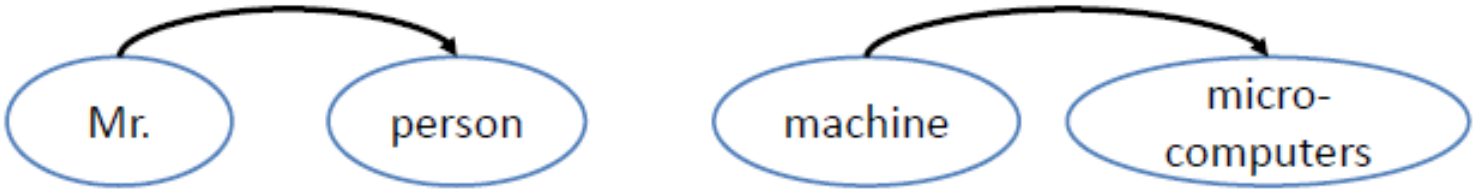
Assumes that important sentences are those that are ‘traversed’ by *strong* chains (Barzilay and Elhadad, 97).

- Strength(C) = length(C) - #DistinctOccurrences(C)
- For each chain, choose the first sentence that is traversed by the chain and that uses a representative set of concepts from that chain.

Lexical Chains(Barzilay97)

Lexical Chain : A chain of semantically related words

Mr. Kenny is the **person** that invented an anesthetic **machine** which uses **micro-computers** to control the rate at which an anesthetic is pumped into the blood...



Lin - set of summarization methods

Sentence order: Sentence order in text gives the importance of the sentences. First sentence highest ranking last sentence lowest ranking.

Title: Sentences containing words in the title get high score.

Term frequency (tf): Open class terms which are frequent in the text are more important than the less frequent. Open class terms are words that change over time.

Position score: The assumption is that certain genres put important sentences in fixed positions. For example. Newspaper articles has most important terms in the 4 first paragraphs.

Lin - set of summarization methods

Query signature: The query of the user affect the summary in the way that the extract will contain these words.

Sentence length: The sentence length implies which sentence is the most important.

Average lexical connectivity: Number terms shared with other sentences. The assumption is that a sentence that share more terms with other sentences is more important.

Numerical data: Sentences containing numerical data obtain boolean value 1 (is scored higher) than the ones without numerical values.

Lin - set of summarization methods

Proper name: Sentences containing proper names will be given higher score

Weekdays and Months: Sentences containing Weekdays and Months will be given higher score

Quotation: Sentences containing quotations might be important for certain questions from user

First sentence: First sentence of each paragraphs are the most important sentences

Lin - set of summarization methods

Decision tree combination function: All the above parameters were put into decision tree and trained on set of texts and manual summarized texts.

Simple combination function: All the above parameter were normalized and put in a combination function with no special weighting.

END