Summarization

Karola Klarl

seminario nel corso

Elaborazione di Linguaggio Naturale

Overview

- Introduction
- Steps of summarization
  - Extraction
  - Interpretation
  - Generation
- Evaluation
- Future

Why summarization?

- Informing
- Decision making
- Time saving

What is a summary?

- Text produced from one or more texts
- That contains a significant portion of information from the original text
- That is no longer than half of the original text

Types of summaries

- Indicative ←→ Informative
- Extract ←→ Abstract
- Generic ←→ Query-oriented
- Single-Doc ←→ Multi-Doc

Paradigms

Information Extraction / NLP

- Top-Down approach
- Query-driven focus
- Query-oriented summaries

- Tries to «understand»
- Need of rules for text analysis at all levels

Information Retrieval / Statistics

- Bottom-Up approach
- Text-driven focus
- Generic summaries

- Operates at lexical level
- Need of large amount of texts

Paradigms

Information Extraction / NLP

+ Higher quality
+ Supports abstracting

- Speed
- Needs to scale up to robust open-domain summarization

Information Retrieval / Statistics

+ Robust

- Lower quality
- Inability to manipulate information at abstract levels

➔ **Combine strength of both paradigms**

Overview

- Introduction
- **Steps of summarization**
  - Extraction
  - Interpretation
  - Generation
- Evaluation
- Future

Steps of summarization

- Extraction
  → Extracts
- (Filtering)
  - Only for Multi-Docs
- Interpretation
  → Templates (unreadable abstract representations)
- Generation
  → Abstracts

Extraction

- General procedure
  - Several independent modules
  - Each module assigns score to each unit of input
  - Combination module combines scores to a single score
  - System returns n highest-scoring units

Methods

- Position-based
- Cue-Phrase
- Word-Frequency
- Cohesion-based
- Discourse-based
- And many more …

Position-based methods

- Lead method
  - Claim: important sentences occur at the beginning (or end) of texts
- OPP (Optimum Position Policy)
  - Claim: important sentences are located at genre-dependent positions

    positions can be determined through training
- Title-based method
  - Claim: words in titles are relevant to summarization

Cue-Phrase method

- Claim 1: important sentences contain «bonus phrases»

  (in this paper we show, significantly, in conclusion)

  non important sentences contain «stigma phrases»

  (hardly, impossible)
- Claim 2: phrases can be detected automatically

Word frequency-based method

- Claim: important sentences contain words that occur frequently

  → Zipf's law distribution

- Generality makes it attractive for further study

Cohesion-based methods

- Claim: important sentences are the highest connected entities in semantic structures
- Classes of approaches
  - Word co-occurrences
  - Local salience and grammatical relations
  - Co-reference
  - Lexical chains
  - Combinations

Discourse-based method

- Claim: coherence structure of a text can be constructed and «centrality» of units reflects their importance
- Coherence structure = tree-like representation

Extraction

- All methods seem to work
- No method performs as well as humans

- No obvious best strategy

Interpretation

- Occurs at conceptual level
- Result = something new, not contained in input
- Need of „world knowledge“, separate from input
  - ➔Really difficult to build domain knowledge
  - ➔Little work so far

Interpretation

- Methods
  - Condensation operators
  - Topic signatures
  - And others …

Condensation operators

- Parse text
- Build terminological representation
- Apply condensation operators
- Build hierarchy of topic descriptions

- Until now no parser/generator has been built!

Topic signatures

- Claim: can approximate topic identification at lexical level using automatically aquired «word families»

- Topic signature is defined by frequency distribution of words related to concept

- Inverse of query expansion in Information Retrieval

Generation

- Level 1: no separate generation

  Produce extracts from input text

- Level 2: simple sentences

  Assemble portions of extracted clauses together

- Level 3: full NL Generation

  Sentence planner: plan content, lenght, theme, order, words , …

  Surface realizer: linearize input grammatically

Overview

- Introduction
- Steps of summarization
  - Extraction
  - Interpretation
  - Generation
- **Evaluation**
- Future

Evaluation

- If you already have a summary
  - Compare the new one to it
  - Choose granularity (clause, sentence, paragraph)
  - Measure similarity of each unit in the new summary to the most similar units in the «gold standard»
  - Measure Precision and Recall

Evaluation

- If you don't have a summary
  - Compression ratio CR = length S / length T
  - Retention ratio RR = info in S / info in T

- RR is measured through Q&A games
  - Shannon game: quantifies information content
  - Question game: test reader's understanding

Overview

- Introduction
- Steps of summarization
  - Extraction
  - Interpretation
  - Generation
- Evaluation
- **Future**

What has to be done …

- Data preparation
  - Collect sets of texts and abstracts
  - Corpora of <text, abstract, extract>
- Types of summaries
  - Determine characteristics for each type
- Extraction
  - New extraction methods
  - Heuristics for method combination

What has to be done …

- Interpretation
  - Investigate types of fusion
  - Create collections of knowledge
  - Study incorporation of user's knowledge in interpretation
- Generation
  - Develop sentence planner rules for dense packing of content into sentences
- Evaluation
  - Better evaluation metrics

Spero che il mio italiano fosse comprensibile ☺

Grazie per l‘attenzione!

References

- The Oxford Handbook of Computational Linguistics, Mitkov R., Oxford: Oxford University Press, pp.583-598 (encyclopedia entry: Automated Text Summarization, Hovy E.H. 2005)
- Automated Text Summarization Tutorial, COLING/ACL '98, by E.Hovy e D.Marcu http://www.isi.edu/~marcu/acl-tutorial.ppt
- Text Summarization Tutorial ACM SIGIR, by D. R. Radev http://www.summarization.com/sigirtutorial2004.ppt
- http://en.wikipedia.org/wiki/Automatic_summarization