Text Summarization

What?

An automatic summary is a text generated by a software, that is coherent and contains a significant amount of relevant information from the source text. Its compression rate τ is less than a third of the length of the original document.

- Produced from one or more documents
- Preserve important information
- Short

Why?

“too much information kills information”

- Professional summarizers
  - Expensive
  - Lacks expertise
- Reduce reading time
- Easier selection of documents
- Improves effectiveness of indexing
- Less biased
- Personalized summaries for QA systems

Summary Categorization

- Extractive
- Abstractive

- Single-document
- Multi-document

- Indicative
- Informative

- Headline summarization
- Ultra-summarization
- Keyword summarization

- Generic
- Query-focused
- Update

Summary Categorization

- Monolingual
- Multi-lingual
- Cross-lingual

- News
- Specialized
- Literary
- Encyclopedic…

- Author
- Expert
- Professional

- Multimedia

Abstractive Summarization

- Understands the text, generate summary (NLG)
- Abstract

- Very difficult

- Compression
- Fusion
- Information Extraction

Extractive Summarization

- Selects sentences from source document
- Extract

- Cohesion
- Coherence
- Unresolved co-references
- Discourse relations

Extractive Summarization

- Intermediate representation
- Scoring sentences
- Selecting summary

Intermediate Representation

- Topic representation
  - VSM, lexical chains, LSA, Bayesian topic models
- Indicator representation
  - sentence length, sentence location, proper nouns, numerical data…
- Graph representation
  - directed forward (backward), undirected

Scoring Methods

- Topic representation
  - ability of a sentence to express topic
- Indicator representation
  - machine learning
- Graph representation
  - stochastic methods

Examples [http://www.sciencedirect.com/science/article/pii/S0957417413002601]

Selecting a summary

- Length constraint

- best n approach
  - Maximal marginal relevance
- Global selection
  - Maximize importance, maximize coherence, minimize redundancy

$$\omega_{\mathrm{MMR}}(s) = \arg\max_{s \in D \setminus \mathrm{Sum}} [\lambda \underbrace{\mathrm{sim}_1(s, Q)}_{\text{Relevance}} - (1-\lambda) \underbrace{\arg\max_{s_\mu \in \mathrm{Sum}} \ \mathrm{sim}_2(s, s_\mu)]}_{\text{Redundancy}}$$

Evaluation

- Manual
- Semi-automatic
  - ROUGE-n
- Automatic

$$\mathrm{ROUGE-n} = \frac{\sum_{\mathrm{n-grams}} \in \{\mathrm{Sum_{can}} \cap \mathrm{Sum_{ref}}\}}{\sum_{\mathrm{n-grams}} \in \mathrm{Sum_{ref}}}$$

- ROUGE-n
  - Lexical level
  - Abbreviations (BEwT-E, PYRAMID)

Frequent Patterns

- Single-document
- Monolingual
- Graph representation
    - Dynamic graph – mimicking reading
    - DGRMiner