Automated Text Summarizer for Turkish

Sefa Kılıç

December 24, 2009

Introduction

- Summary is text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that {Radev, Hovy, McKeown, 2002}.
- Two contrary aims:
  - Summary should be short as much as possible.
  - Summary should contain the information in the original text(s) as much as possible.

Introduction (cont.)

Summaries can be classified as

- abstractive or extractive
- one document or multi document
- query-oriented or generic

Introduction (Cont.)

Summaries can be classified as

- abstractive or **extractive**
- **one document** or multi document
- query-oriented or **generic**

Outline

- Related Work
- Summarization using features
- Summarization using Latent Semantic Analysis
- Evaluation

Related Work

- H. P. Luhn. The automatic creation of literature abstracts. (1958)
- Chin-Yew Lin. Training a selection function for extraction. (1999)
- Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. (2001)
- Chin-Yew Lin and E. Hovy. Manual and automatic evaluation of summaries (2002)
- Chin-Yew Lin, Rouge: A package for automatic evaluation of summaries (2004)

Summaries Using Features

Features are used for ranking sentences. {Chin-Yew Lin, 1999}

$$Score(S) = \sum_i w_i Score_{f_i}(S)$$

where w_i indicates the weight of the feature f_i to the overall score. Following features are used in the calculation of overall scores for each sentence.

- Sentence position
- Title
- Term frequency
- Average lexical connectivity
- Numerical data
- Proper name
- Pronoun and adjective
- Weekday and month
- Quotation

Summaries Using Latent Semantic Analysis

- {Y. Gong, X. Liu, 2001}
- Create a terms by sentences matrix $\mathbf{A} = [A_1 A_2 \ldots A_n]$ with each column vector A_i representing the weighted term-frequency vector of sentence i.
- $\mathbf{A}$ is $m \times n$ matrix where m is the number of terms and n is the number of sentences.
- For $\mathbf{A}$, where $m \geq n$ without loss of generality, SVD of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$$

  - $\mathbf{U} = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors.
  - $\Sigma = diag(\sigma_1, \sigma_2 \ldots \sigma_n)$ is $n \times n$ diagonal matrix whose diagonal elements are non-negative singular values sorted in descending order.
  - $\mathbf{V} = [v_{ij}]$ is an $n \times n$ orthonormal matrix whose columns are called right singular vectors.

Summaries Using Latent Semantic Analysis (Cont.)

- SVD derives the latent semantic structure from document.
- Each column vector i of $\mathbf{A}$ is mapped to $\Psi_i = [v_{i1} v_{i2} \ldots v_{ir}]^T$ of matrix $\mathbf{V}^T$.
- k'th singular vector represents k'th important concept/topic.
- Because all singular vectors are independent from each other, the sentences selected by this method contain the minimum redundancy.
- SVD is capable of capturing and modeling interrelationships among terms.

Summaries Using Latent Semantic Analysis (Cont.)

Gong and Liu propose the method in six steps:

1. Decompose the document D into individual sentences and use these sentences to form the candidate sentence set S, and set $k = 1$.
2. Construct terms by sentences matrix A for the document D.
3. Perform the SVD on A to obtain singular value matrix Σ, and the right singular vector matrix V^T. In the singular vector space, each sentence i is represented by the column vector $\Psi_i = [v_{i1} v_{i2} \dots v_{ir}]^T$ of V^T.
4. Select the k-th right singular vector from the matrix V^T.
5. Select the sentence which has the largest index value with the k-th right singular vector, and include it in the summary.
6. If k reaches the predefined number, terminate the operation; increment k by one, and go to Step 4.

Data Corpus

- News articles constructed in {Cigir, Kutlu, Cicekli 2009}.
- News articles from Turkish news portals.

Performance Evaluation

- Recall (R), precision (P) and F-measure (F) are used.

$$R = \frac{|S_{man} \cap S_{auto}|}{|S_{man}|} \qquad P = \frac{|S_{man} \cap S_{auto}|}{|S_{auto}|} \qquad F = \frac{2PR}{P+R}$$

  - S_{man} is the set of sentences selected by human evaluator(s).
  - S_{auto} is the set of sentences selected by the system.
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation {Chin-Yew Lin, 2004}
  - ROUGE-N: N-gram based co-occurrence statistics
  - ROUGE-L: LCS based statistics

References

- {Radev, Hovy, McKeown, 2002} D. R. Radev, E. H. Hovy, K. McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399-408, 2002.
- {Chin-Yew Lin, 1999} Chin-Yew Lin. Training a selection function for extraction. In CIKM'99: *Proceedings of the eighth conference of Information and knowledge management*, pages 55-62, New York, NY, USA, 1999. ACM.
- {Y. Gong, X. Liu, 2001} Y. Gong and X.Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR'01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval,* pages 19-25, New York, NY, USA, 2001. ACM.
- {Cigir, Kutlu, Cicekli 2009} C. Cigir, M. Kutlu, I. Cicekli. Generic text summarization for Turkish. In *Proceedings of the 24th International Symposium on Computer and Information Sciences,* 2009
- {Chin-Yew Lin, 2004} Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out.* page 10, 2004