Latent Dirichlet Allocation for Comparative Summarization

Michal Campr[1]

1 Introduction

This paper explores the possibility of using Latent Dirichlet Allocation (LDA) for multi-document comparative summarization which is quite a recent area of research. The aim is to find some latent information about the input documents and summarize factual differences between them. These differences are then represented by the most characteristic sentences which form the resulting summaries.

2 Latent Dirichlet Allocation

LDA, published by Blei (2003), is a model which breaks down the collection of documents (the importance of a document for the document set is denoted as $P(D)$) into topics by representing the document as a mixture of topics with a probability distribution representing the importance of *j-th* topic given a document (denoted as $P(T_j|D)$). The topics are represented as a mixture of words with a probability representing the importance of the *i-th* word for the *j-th* topic (denoted as $P(W_i|T_j)$). The most often used method for obtaining the topic and word probabilities is the Gibbs sampling method Phan (2013).

3 The algorithm

The first step is loading the input data from two document sets *A* and *B*, removing stopwords (e.g. prepositions or conjunctions) and performing term lemmatization to improve the semantic value of the documents. By running the Gibbs sampler, we obtain the word-topic distributions for each document set and store them in matrices T_A (topic-word) for the document set *A* and T_B for *B*. We can then compute topic-sentence matrices U_A and U_B with sentence probabilities:

$$P(S_r|T_j) = \frac{\sum_{W_i \in S_r} P(W_i|T_j) * P(T_j|D_r)}{length(S_r)^l}, \tag{1}$$

where $l \in < 0, 1 >$ is an optional parameter to configure the handicap of long sentences. Next step covers the creation of two diagonal matrices SIM_A and SIM_B which contain the information about similarities of topics from both sets. This is accomplished in two steps:

1. $T_A = [T_{A1}, T_{A2}, ..., T_{An}]^T, T_B = [T_{B1}, T_{B2}, ..., T_{Bn}]^T$, where T_{Ai} and T_{Bi} are row vectors representing topics and *n is the number of topics. For each* T_{Ai} *find* red_i *(redundancy of i-th topic) by computing the largest cosine similarity between* T_{Ai} *and* T_{Bj}*, where* $j \in < 1..n >$ *and storing value* $1 - red_i$ *representing the novelty of i-th topic into matrix* SIM_A*.*

[1] PhD student on the Department of Computer Science and Engineering, FAV, University of West Bohemia, e-mail: mcampr@kiv.zcu.cz

2. *For each T_{Bi} find red_i (redundancy of i-th topic) by computing the largest cosine similarity between T_{Bi} and T_{Aj}, where $j \in < 1..n >$ and storing value $1 - red_i$ representing the novelty of i-th topic to matrix SIM_B.*

Finally, we create matrices $F_A = SIM_A * U_A$ and $F_B = SIM_B * U_B$ and from these, sentences with the best score are selected and included in the summary.

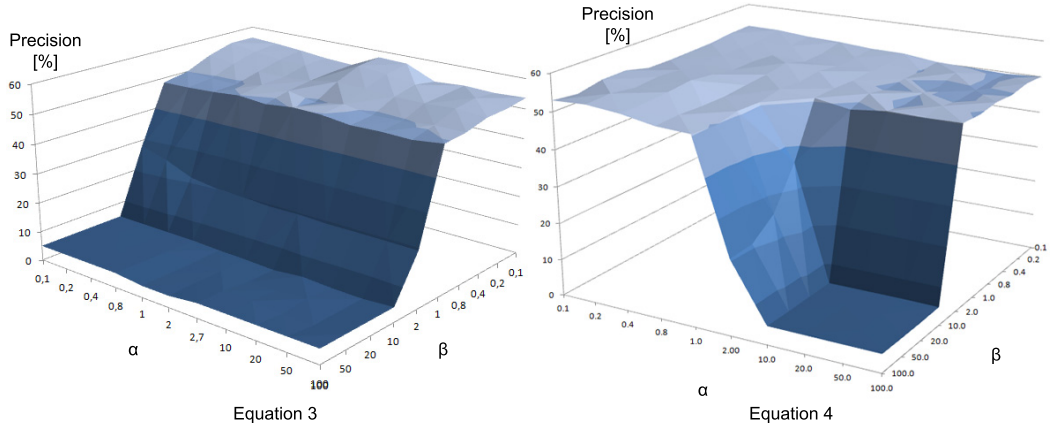

Figure 1: Average precision depending on parameters α and β for variations of equation 1

4 Evaluation

Due to the lack of unified testing data for the task of comparative summarization, we created our own data set for evaluation. We utilized dataset from TAC 2011 conference which consists of 100 news articles, divided into 10 topics. We created pairs of sets of documents by combining different topics to simulate documents which have something in common, but also some differences.

The last issue of the proposed method is how to set the parameters for the Gibbs sampler to get the best results. We tested different values for both parameters α and β. Values varied from 0 to 100, and we computed the average precision. The result is on the Figure 1. The best average precision value we were able to achieve was $57,74\%$.

References

DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

Haghighi, Aria and Vanderwende, Lucy. Exploring content models for multi-document summarization. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*

Xuan-Hieu Phan, Cam-Tu Nguyen. *http://jgibblda.sourceforge.net/.*

Tiedan Zhu and Kan Li. The Similarity Measure Based on LDA for Automatic Summarization. Procedia Engineering*, 2012.*