Comparative Summarization via Latent Dirichlet Allocation

Michal Campr and Karel Jezek

Department of Computer Science and Engineering, FAV, University of West Bohemia,
11 February 2013, 301 00, Plzen, Czech Republic
{mcampr, jezek_ka}@kiv.zcu.cz

Abstract. This paper aims to explore the possibility of using Latent Dirichlet Allocation (LDA) for multi-document comparative summarization which detects the main differences in documents. The first two sections of this paper focus on the definition of comparative summarization and a brief explanation of using the LDA topic model in this context. In the last three sections, our novel method for multi-document comparative summarization using LDA is presented and also its results are compared with the results of a similar method based on Latent Semantic Analysis.

Keywords: comparative summarization, latent dirichlet allocation, latent semantic analysis, topic model

1 Comparative summarization

With the continuing grow of the internet as a source of information, the need for data compression is obvious. This necessity does not apply only to audio or video, but also to textual data (i.e. text summarization). As the amount of textual data grows, the probability of duplicate documents, or documents with very similar features, arises. This is the main problem that we are focusing on in this particular paper and we explore the possibility of utilising the Latent Dirichlet Allocation (LDA) topic model. Comparative summarization is quite a recent area of research and several methods have already been explored. The purpose of these methods is to find some latent information about the input documents and find factual differences between them. These differences are then represented by the most characteristic sentences which form the resulting summaries.

2 Text summarization via LDA

Latent Dirichlet Allocation has already been utilized in several methods, but to our knowledge it has not yet been used in the context of comparative summarization. The closest problem already addressed is the so called update summarization. It aims to search for information, which newly arise in a series of

documents about the same topic. The assumption is that the user is familiar with one document and would like to know what information are additional in another document. We have investigated the already published methods for basic and update summarization using LDA to learn the possibilities of comparing two sets of documents so that we can utilise the best practises to address the problem of comparative summarization.

2.1 Basic summarization via LDA

Latent Dirichlet Allocation (LDA) [4] can be basically viewed as a model which breaks down the collection of documents (the importance of document B for the document set is denoted as $P(D_B)$) into topics by representing the document as a mixture of topics with a probability distribution representing the importance of *j-th* topic for document B (denoted as $P(T_j|D_B)$). 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 $P(W_i|T_j)$). This model has already been used for basic summarization in several papers. The topic and word probabilities are in each of the below mentioned methods obtained using the Gibbs sampling method [1]. These summarization methods are briefly described in the following paragraphs. In order to shorten the explanations, only some interesting ideas and explanations (for the purpose of this paper)are mentioned.

The paper [3] has presented new algorithms for scoring sentences based on LDA probability distributions. The basic idea is computing the probability of the r-th sentence from probabilities of words and topics (depending on used algorithm):

$$P(S_r|T_j) = \prod_{W_i \in S_r} P(W_i|T_j) * P(T_j|D_B) * P(D_B) \tag{1}$$

or

$$P(S_r|T_j) = \frac{\sum_{W_i \in S_r} P(W_i|T_j) * P(T_j|D_B) * P(D_B)}{length(S_r)} \tag{2}$$

After obtaining the probabilities $P(S_r|T_j)$, i.e. the probabilities of *r-th* sentence belonging to the *j-th* topic, the selection of the most significant sentences can begin. The process is finished when the number of sentences reaches a predefined amount.

The other paper dealing with LDA-based summarization is [2]. The idea is to combine the LDA topic model and Latent Semantic Analysis (LSA) to reduce the information content in sentences by their representation as orthogonal vectors in a latent semantic space. At first, the LDA probability distributions of topics and words are obtained. After that, for each topic T_j, a term-sentence matrix is created and then the Singular Value Decomposition (SVD) is applied to each of them. The result of the SVD are three new matrices U, Σ and V^T, from which only the third one is utilised. This matrix contains the so called right singular vectors, which basically map topics to sentences. After obtaining the sentence

probabilities, the process of selecting sentences with the best score can run until the predefined summary length is reached.

The paper [8] presents two algorithms for summarization and most importantly a new sentence similarity measure based on LDA. Instead of representing a sentence as a sparse vector using tf-idf, the idea is to use the LDA topic model to represent words and sentences as vectors of topic probabilities. The sentence vector is calculated as an average value of topic vectors of all words in the given sentence. Using this representation, it is a simple matter to measure the similarity between any two vectors using cosine similarity. The summarization algorithms are then based on selecting the best candidate sentence which also has the lowest redundancy with the existing summary until the summary length is reached.

2.2 Update summarization via LDA

The update summarization is the closest problem to ours, so we explored the used methods of comparing LDA topics. The following paragraphs describe methods of update summarization that have been already published and evaluated.

In the paper [6] a novel update summarization framework was proposed. The topics were extracted from two sets of documents A and B by the means of LDA topic model. The topics were assigned into four different categories:

- emerging – topics that newly arise in B
- activating – topics in both set, but with more emphasis in B
- non-activating – topics in both sets, but not too much discussed in B
- perishing – topics only in A

The correlations between old and new topics were then identified with the use of Pearson product-moment correlation. A novel algorithm (CorrRank) was also developed for ranking sentences with topic correlation so that the best ranked sentences can be iteratively added to the resulting summary.

The method proposed in the paper [5] is derived from TopicSum presented in [7] and the topic model of input documents is restricted to only two topics for each document set. The idea is that one topic in each document contains all the already known facts and the second topic contains all the new information that we want to extract.

3 Comparative summarization via LDA

This section will thoroughly describe our novel method for comparative summarization using LDA topic model. Our idea is to use this topic model to represent the documents, compare these topics and select the most significant sentences from the most diverse topics, to form a summary.

The first step is to load the input data from two document sets A and B. The important thing here is that from the perspective of LDA, we treat every

sentence as one document. When we have all the sentences from both sets loaded, we can estimate the LDA parameters (the exact reason will be discussed in the last section of this paper) as follows:

- summaryLength = 10sentences
- numberOfTopics = $\sqrt{\text{numberOfSentences}}$
- numberOfIterations = 3000
- $\alpha = 50/\text{numberOfTopics}$
- $\beta = 200/\text{numberOfWords}$

Before we run the Gibbs sampler (we used the implementation JGibbLDA from [1]) to obtain the LDA topics, we have to remove the stop-words and perform term lemmatization. This way we are sure that there are no words that carry no useful information. With the parameters set and input text prepared, we can 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, where row vectors represent topics and column vectors represent words. A very important aspect of writing the distributions into matrices is to ensure that both of them have the same dimensions, i.e. to work as well with the words that appear only in one set and including them also in the second matrix (with zero probability). After this, we can compute topic-sentence matrices U_A and U_B with sentence probabilities (we experimented with two equations):

$$P(S_r|T_j) = \frac{\sum_{W_i \in S_r} P(W_i|T_j)}{length(S_r)^l}, \tag{3}$$

or

$$P(S_r|T_j) = \frac{\sum_{W_i \in S_r} P(W_i|T_j) * P(T_j|D_r)}{length(S_r)^l}, \tag{4}$$

where $l \in <0, 1>$ is an optional parameter to configure the handicap of long sentences. The row vectors represent topics and the columns are 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.
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 $F_B = SIM_B * U_B$ combining the probabilities of sentences with the novelty of topics. From these matrices, it is a simple matter to find sentences with the best score and including them

in the summary. For better results, it is essential to compare the candidate sentence with already selected sentences to avoid information redundancy (the comparison is also achieved via the cosine similarity). If a sentence is selected, the relevant vector in F_A or F_B is set to *0* in order to remove the information from the matrix. The final result consists of two independent summaries of predefined length, each of which depicts the most significant information, which are specific for one of the compared document set exclusively.

4 Evaluation

Due to the lack of unified testing data for the task of comparative summarization, we had to create our own data set for evaluation. We have utilised data from TAC 2011 conference to find out if the proposed method brings the expected results. The available data consist of 100 news articles in total, divided into 10 topics, 10 articles each. With these articles, we have created pairs of sets of documents by combining different topics (Figure 1). In every pair, there is one identical topic present in both sets and one topic for each of the sets that are different. This has a simple purpose: to simulate two sets of documents which have something in common, but also some differences. This setup allows us also to easily compute the precision of selecting sentences because we know which sentences we want the algorithm to select. The reason for the use of TAC 2011 dataset is also the fact, that there are three human-created summaries for each of the 10 topics. This allows us to further evaluate our method with the ROUGE toolkit. However, the ROUGE based evaluation is not included in this paper, because it is not yet complete.

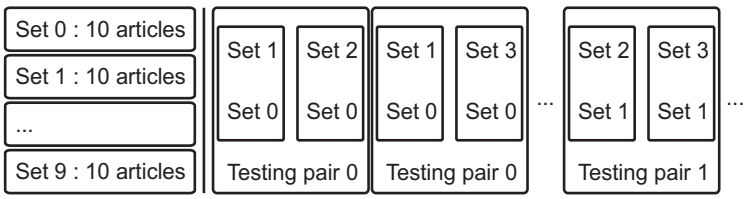

Fig. 1. Creating testing data-sets

Another problem we encountered was how to compare two vectors to gain the best results. We tried two possibilities: cosine similarity and Pearson correlation (as was mentioned in [6]). From these two options, cosine similarity gave better results and comes out as a better choice, even if the precision was only higher in the order of tenths percent.

The last issue of the proposed method is how to set the parameters for the Gibbs sampler to get the best LDA distributions. We have tested our method

on 11 values for both parameters α and β, including values recommended in Section 4 (those depending on the number of sentences or words). Parameter values varied from 0 to 100, and we computed the average precision. The result is on the Figure 2. As can be seen, the α parameter has only a little impact on the precision if the equation 3 is used. On the other hand, for the equation 4, the impact on precision is practically the same as for the β parameter. At the end, the best overall average precision value we were able to achieve was $57,74\%$.

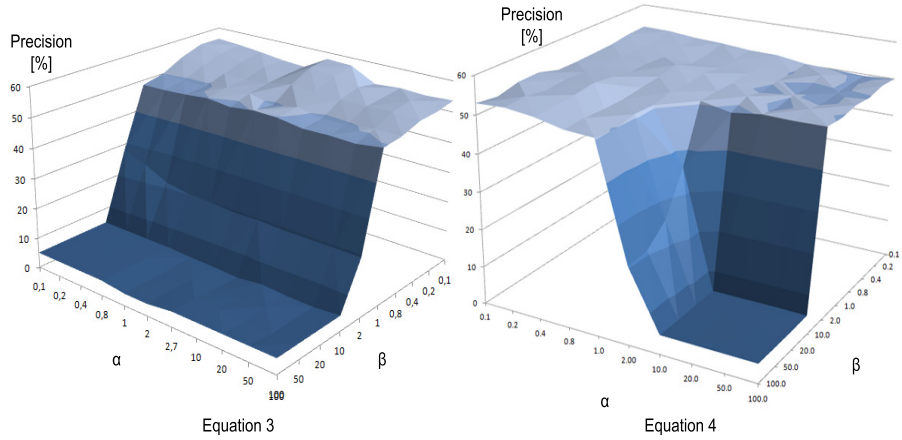

Fig. 2. Average precision depending on parameters α and β for equations 3 and 4

5 Conclusion

In our previous work, we developed a similar method for comparative summarization using Latent Semantic Analysis. In this case, the average precision values were in the range from $61,23\%$ to $98,44\%$ for different configurations of the algorithm. Although the LDA provides more intuitive topic model, it has evidently much lower precision values for any case of given parameters and thus the LSA comes out as a better choice for comparative summarization. The last step in evaluating these two methods is via the ROUGE toolkit, which we are working on right now.

Our future work resides still in the area of comparative summarization, but we would like to explore the possibilities of including sentiment analysis in the process of topic comparison in order to widen the area of usability.

Acknowledgements

The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

References

[1] Xuan-Hieu Phan, Cam-Tu Nguyen. `http://jgibblda.sourceforge.net/`.

[2] Arora, Rachit and Ravindran, Balaraman. Latent dirichlet allocation and singular value decomposition based multi-document summarization. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. ICDM'08. Eighth*, pages 713–718, 978-0-7695-3502-9.

[3] Arora, Rachit and Ravindran, Balaraman. Latent dirichlet allocation based multi-document summarization. *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 91–97, Singapore, 978-1-60558-196-5.

[4] DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, pages 993–1022, 2003.

[5] Delort, Jean-Yves and Alfonseca, Enrique. DualSum: a Topic-Model based approach for update summarization. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223, 2012.

[6] Lei Huang and Yanxiang He. CorrRank: update summarization based on topic correlation analysis. *In proceedings of 6th International Conference on Intelligent Computing*, pages 641–648, 2010.

[7] 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*, pages 362–370, 978-1-932432-41-1

[8] Tiedan Zhu and Kan Li. The Similarity Measure Based on LDA for Automatic Summarization. *Procedia Engineering*, pages 2944–2949, January 2012.