Ranking topic model based automatic summary: a win-win situation

Anonymous Author(s)
Affiliation
Address
email

Abstract

To alleviate the problem that learned topic distributions of most topic models have no orders, a ranking topic model based on correlated topic model is proposed in this paper. Two new features, topic correlation and topic quality, are used to rank topics. Automatic summary is adopted as extended task to indirectly measure the effectiveness of propose algorithm. Experimental results show that automatic summary evaluation metric can indeed measure the performance of different topic model based summary algorithms. Meanwhile, ranking topic model based automatic summary algorithm exhibits superior performance compared with other topic model based algorithms. Hence, ranking topic model and automatic summary form a win-win pair as they benefit from each other.

1 Introduction

Topics are features with rich semantics, which have been used in various tasks, such as automatic summary. In topic model based automatic summary, learned topics are used to select prominent sentences. Previous works have shown that automatic summary benefits from topic features. But how can topic model benefit from summary, and is there a way for these two to benefit from each other and have a win-win situation?

A large portion of automatic summary study mainly focuses on extractive summary. Selecting prominent sentences that cover most of the key points of the document is the most important task.[1]. Using topic as feature to select sentences requires topic to be *semantically* discriminative. But as far as we know, learned topic distributions of most topic models couldn't distinguish from each other. A ranking topic model called CorrRank is proposed to solve the topic ranking problem.

CorrRank solely uses the learned model of Correlated Topic Models (CTM) [2] to re-order topic distributions. Specifically, topic correlation and topic quality are learned as new features from the result of CTM. Combined together, these two features assign a ranking score to each topic. With re-ordered topics by their ranking score as feature, semantically prominent sentences containing terms from high-order topics can be easily selected.

Meanwhile, there is no well defined and recognized metrics to evaluate the performance of ranking topic model. Since topics, ranked or not, are the only changing factor, evaluating automatic summary algorithm using existing metrics can indirectly measure the effectiveness of ranking topic model. **In this way, automatic summary obtains a more semantic-rich feature; ranking topic model gets an external evaluation method.**

The following paper is organized as follows: in chapter 2 we review other works concerning ranking methods in topic models. In chapter 3, we formally present CorrRank algorithm. In chapter 4, we conduct various experiments to test our model and show how ranking topic model and automatic summary benefit from each other. We conclude our work and point out the future work in the end.

2 Related works

AlSumait et al. proposed the problem of re-ranking topic distributions according to their importance [3]. They defined important topics and irrelevant topics in three different manners, and then used weighted scores derived from three manners to rank topic distributions accordingly. They were the first to raise the problem that an ordered topic list with ranking is necessary for the model. Lau et al. proposed methods to select the appropriate words to represent topics, this can be seen as re-ranking terms in each topic [4]. Both work have shown that ordered topics can increase the usability of topic. Furthermore, ranking schemes have been incorporated in extractive summary study [5, 6]. Meanwhile, previous studies have shown that topic features can increase the performance of automatic summary [7, 8]. It seems natural to combine these two features together. But to our knowledge, there is little work investigating whether organized topic features can benefit automatic summary.

3 CorrRank

3.1 Topic correlation

In CTM, logistic normal distribution is used to represent the prior knowledge. Logistic normal distribution has two parameters, one is the mean μ, and the other is the covariance Σ. In Σ, each column represent a topic. For each topic, we calculate its connection degree with other topic with equation 1:

$$\textbf{Topic Correlation}: TC_k = \sum_{i,j=1,i\neq j}^{K-1} \sigma_{ij} \quad k \in \{1, 2, \cdots, K-1\} \tag{1}$$

σ_{ij} is the element of matrix Σ, k is the topic index. The connection degree between topics reflects the popularity between topics, topic with high connection degree means that the semantic of this topic is closer to other topics, topic with low connection degree means that the topic is more isolated.

Although TC_k reflects the relationships between topics, but it is highly biased. It only takes the corpus level information into consideration, but neglects the actual meaning of terms. For example, if a topic that contains a lot of popular but meaningless terms, it will be quite popular but be of less use to users. For this reason, we have to take document level and topic level information into account.

3.2 Topic Quality

Topic-document frequency is adopted to measure the topic popularity on document level, which didn't reflect any correlation between topics. The higher this value, the more a topic prevails in the corpus. This reflects the proportion of one topic in a corpus. Topic-document frequency is calculated in equation 2:

$$\textbf{Document Frequency}: DF_k = \frac{d_k}{D} \quad k \in \{1, 2, \cdots, K-1\} \tag{2}$$

d_k is the number of document that contain the kth topic, D is the total number of the documents in corpus.

Topic significance is used to balance the topic correlation. As aforementioned, popular topics may contain popular but meaningless terms, [3] called them "junk" words. On the other hand, a small set of words that have genuine meaning are called "salient" words. In one document, there are always large number of junk words and small set of salient words. A topic is expected to have a unique character, thus if a topic is closer to the empirical distribution of words in a document, the less uniqueness this topic possesses, hence less significance. Significance of a topic is defined by the distance between the topic and the empirical distribution of the document which contains this topic. The empirical distribution of each document is defined as the probability of each term contains in the document. The topic significance score is defined in equation 3:

$$\textbf{Topic Significance}: TS_k = -\sum_{i=1}^{D_k} KL\left(\phi_k \parallel p_i\right) \quad k \in \{1, 2, \cdots, K-1\} \tag{3}$$

ts denotes topic significance score, K is topic number, D_k is the number of document which contains topic k , ϕ_k is the k-th topic distribution. KL divergence is used to calculate the difference between the topic distribution and the empirical distribution.

In all, document frequency and topic significance are combined to assess the topic quality.

$$\textbf{Topic Quality}: TQ_k = DF_k \times TS_k \tag{4}$$

Here, document frequency functions like a normalizer to prevent topics that prevail the corpus having too big weight.

The final score is then combined to rank the topic distribution, we use a parameter α to control the balance between topic correlation and topic quality. The topic ranking score is then defined in equation 5:

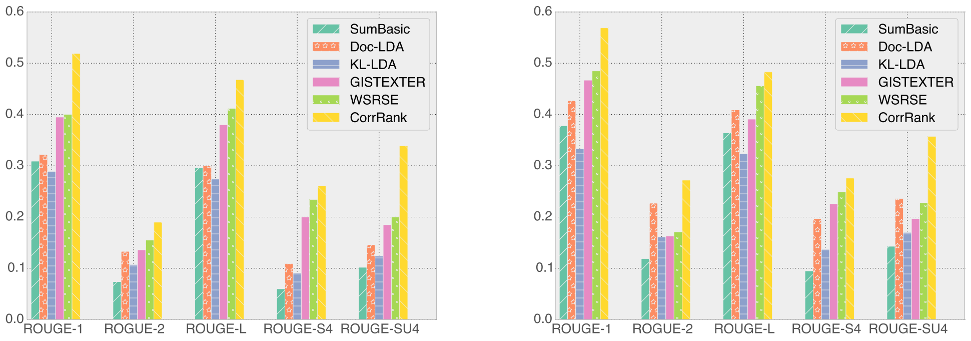
$$\textbf{RankScore} = \alpha \cdot TC_k + (1 - \alpha) \cdot TQ_k \tag{5}$$

We set α value to 0.6, which leans a little heavy over topic correlation part. The parameter is meant to be set manually by user to emphasis on either topic significance or document frequency.

RankScore is used to re-order topics, ordered topics are then used as features to select semantically prominent sentences.

4 Experiments and discussions

Well acknowledged metric in automatic summary ROUGE [9] is adopted to evaluate the performance of automatic summary algorithms and the effectiveness of ranking topic model. CorrRank is compared with SumBasic, Doc-LDA, KL-LDA. GISTEXTER[10] and WSRSE[11] that uses word frequency as features is adopted as baseline. If CorrRank based algorithm performs better, then this proves the effectiveness of our ranking scheme. We denote CorrRank based multi-document summary as CorrRank. We carried out experiments on DUC 2002 corpus dataset, we examined summary of length 200 and 400. Experimental results are shown in figure 1.



(a) 200 words summary

(b) 400 words summary

Figure 1: Rouge scores of different summary algorithms

The experimental results reveal the superiority of the ranked topic feature. In both 200 words summary(figure 1a)) and 400 words summary(figure 1b), CorrRank outperform all other comparison algorithm on all five ROUGE scores.

Compared with other topic feature based automatic summary algorithms, CorrRank excels for two reasons. First, Doc-LDA and KL-LDA only considers document frequency and topic significance separately, CorrRank combines these two features into topic quality feature. Second, CorrRank emphasizes on topic correlation, which set more influencing topics on higher rank. These two factors help to select more semantically prominent sentences.

Comparison between word frequency feature based and topic feature based automatic summary algorithms reveals an interesting result. Other than CorrRank, the other topic feature based algorithm perform worse in most measures in both 200 and 400 words summary. Only Doc-LDA overpasses word frequency feature based algorithm in three measures in 400 words summary. This result reveals that topic feature sometimes are not very stable. CorrRank can always choose the most semantically discriminative topics, which makes selected topic features more robust and constant.

5 Conclusion

In this paper, we proposed a ranking topic model called CorrRank. CorrRank ranks the learned topic distributions of CTM. Ranked topics exhibit superior character as features to select semantically prominent sentences in automatic summary task. Meanwhile, recognized evaluation metric in automatic summary such as ROUGE, is borrowed to evaluate the effectiveness of ranking topic model. Experimental results show that ordered topic discovery and automatic summary can indeed benefit from each other. Since topic discovery and summary are both tasks highly related to human being's abstract thinking process, direct feedback and evaluations are the most appropriate and highly desired informations in these two tasks. In the future, we intend to use crowd-sourcing techniques to design a learning system, utilizing the proposed ranking topic model based automatic summary algorithm to help Chinese students learn English comprehension test in standard English tests. While helping students learn English, the system can collect participants' actual responses that can be used to further improve the algorithm.

References

[1] Dipanjan Das and André FT Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195, 2007.

[2] David M. Blei and John D Lafferty. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18*, 2006.

[3] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of LDA generative models. *Machine Learning and Knowledge Discovery in Databases*, pages 67–82, 2009.

[4] Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. Best topic word selection for topic labelling. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 605–613. Association for Computational Linguistics, 2010.

[5] G Erkan and D R Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

[6] Xiaojin Zhu, Andrew B Goldberg, Jurgen Van Gael, and David Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, pages 97–104, 2007.

[7] Hal Daumé, III, and Daniel Marcu. Bayesian query-focused summarization. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2006.

[8] Sanda Harabagiu and Finley Lacatusu. Using topic themes for multi-document summarization. *ACM Transactions on Information Systems*, 28(3):1–47, June 2010.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[10] Sanda M Harabagiu and Finley Lacatusu. Generating single and multi-document summaries with gistexter. In *Document Understanding Conferences*, 2002.

[11] Hans Van Halteren. Writing style recognition and sentence extraction. In *U. Hahn and D. Harman (Eds.), Proceedings of the workshop on automatic summarization*, pages 66–70. Citeseer, 2002.