Topic-driven Multi-Document Summarization

Hongling Wang   Guodong Zhou
Jiangsu Provincial Key Lab for Computer Information Processing Technology
School of Computer Science and Technology
Soochow University, Suzhou, China 215006
{hlwang, gdzhou}@suda.edu.cn

Abstract—**This paper presents a topic-driven framework for generating a generic summary from multi-documents. Our approach is based on the intuition that, from the statistical point of view, the summary's probability distribution over the topics should be consistent with the multi-documents' probability distribution over the inherent topics. Here, the topics are defined as weighted "bag-of-words" and derived by Latent Dirichlet Allocation from a collection of documents, either the given multi-documents or a related large-scale corpus. In this sense, we could represent various kinds of text units, such as word, sentence, summary, document and multi-documents, using a single vector space model via their corresponding probability distributions over the derived topics. Therefore, we are able to extract a sentence or summary by calculating the similarity between a sentence/summary and the given multi-documents via their topic probability distributions. In particular, we propose two methods in similarity measurement: the static method and the dynamic method. While the former is employed to detect the salience of information in a static way, the later further controls redundancy in a dynamic way. In addition, we integrate various popular features to improve the performance. Evaluation on the TAC 2008 update summarization task shows encouraging results.**

Keywords- Multi-document Summarization; Topic Modeling; Latent Dirichlet Allocation; Static Method; Dynamic Method

I. INTRODUCTION

Multi-document summarization (MDS), a process of producing a single summary from a set of documents, is still deemed one of the hardest tasks in NLP.

In general, a set of documents contains several topics rather than a single one even though it has a central topic. Summaries created by human beings tend to cover several topics to give the readers an overall idea about the given multi-documents. This fact indicates that the summary's probability distribution over the topics should be consistent with the multi-documents' probability distribution over the inherent topics from the statistical point of view.

This paper proposes a topic-driven framework for generating a generic summary from multi-documents. We assume that the given multi-documents contain a central topic as the main property as well as other topics which support around the central one. Together, the central topic and the other ones form a topic probability distribution for the set of documents. Here, a topic is defined as weighted "bag-of-words" and a topic model is employed to derive the inherent topics from a set of documents, either the given multi-documents or a related large-scale corpus. In this way, we can represent various kinds of text units, such as word, sentence, summary, document and multi-documents, using a single vector space model via their corresponding probability distributions over the derived topics. Following this way, sentences can be selected incrementally from the given multi-documents to form a generic summary by simply calculating the similarity between a sentence/summary and the given multi-documents via their topic probability distributions. In particular, a static method is employed to detect the salience of information in a static way and a dynamic method is proposed to further control redundancy in a dynamic way.

The rest of this paper is as follows. Section II gives an overview of related work. Section III gives a brief introduction to topic modeling, in particular LDA. Section IV presents our topic-driven MDS framework in details. Section V presents experimental results. Finally, Section VI draws a conclusion.

II. RELATED WORK

Although topic-driven multi-document summarization has been explored to a certain extent in recent years, our approach is different from those in the literature.

Similar to our work, Arora and Ravindran (2008), Haghighi and Vanderwende (2009), Wang et al. (2009), and Bhandari et al. (2008) use a topic model to extract a summary. In particular, Arora and Ravindran (2008) first use LDA to find the different topics in the documents and then use SVD to find the sentences that best represent the topics. Haghighi and Vanderwende (2009) utilize a hierarchical LDA-style model to represent a topic as a hierarchy of topic vocabulary distributions. Wang et al. (2009) propose a Bayesian topic model at the sentence level by making use of both the term-document and term-sentence associations. Bhandari et al. (2008) use PLSI, another popular topic modeling tool, to divide a document into several topics, and then extract the sentences from the highest-ranked topic as the basic summary and gradually combines the sentences from different topics to form the final summary. Although this approach is successful in single document summarization, it may not be readily applied to multi-document summarization. The reason is that it is difficult for this approach to cope with the huge redundancy in the given multi-documents. For example, in the TAC 2008 update summarization task, the participants are asked to provide a 100-word summary given 10 newswire articles. As we know, long sentences are quite common in newswire articles. Therefore, a 100-word summary is too short to cover all of topics in the set of documents. In our approach, we take into consideration the topic probability distribution of a sentence in the document collection, rather than the single probability that the sentence occurs in one major topic.

978-0-7695-4288-1/10 $26.00 © 2010 IEEE
DOI 10.1109/IALP.2010.26

This paper proposes a topic driven MDS framework which well integerate sentence selection and redundancy removal.

III. TOPIC MODELING

Among various topic models, Latent Dirichlet Allocation (LDA, Blei et al., 2003) has drawn most attention recently in the NLP community and has been applied successfully in topic detection. In this paper, we use LDA to capture the topics in the documents. In particular, we employ the LDA toolkit[1] developed by Phan and Nguyen in our multi-document summarization system.

In this paper, We investigate two ways to apply LDA. One is to derive a single topic model using all documents in the entire corpus of given multi-document sets (called **SingleModel**). Another way is to derive a topic model for each set of given multi-documents (called **MultiModel**). Here, both LDA models use the following default parameters: $\alpha = 50/K,\ \beta = 0.1$, where K stands for the number of topics. In addition, SingleModel sets the number of topics K to 50 and the number of iterations to 3000, while MultiModel sets the number of topics K to 10 and the number of iterations to 2000.

IV. TOPIC-DRIVEN MULTI-DOCUMENT SUMMARIZATION

Our topic-driven multi-document summarization framework consists of three key components: sentence selection, summary generation and summary compression.

A. Sentence selection

Our sentence selection method is based on the intuition that the summary's topic probability distribution should be similar to the multi-documents' topic probability distribution. That is, for a sentence in the summary, its topic probability distribution should be similar to the documents' topic probability distribution, and, for the summary as a whole, its topic probability distribution should be consistent with the documents' topic probability distribution.

In this paper, we mainly investigate Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) to measure the distance between two probability distributions as follows:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Since the KL divergence is asymmetric, we apply the following KL divergence-based symmetric measure:

$$KL(P, Q) = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)$$

where P represents the document's probability distribution over topics, i.e.

$$P = (P(T_1 \mid D_i), P(T_2 \mid D_i), \dots, P(T_k \mid D_i))$$

where k is the number of topics, $P(T_j|D_b)$ stands for the probability of being topic T_j given document D_b. Q represents the sentence's probability distribution over topics, i.e.

$$Q = (P(T_1 \mid S_i), P(T_2 \mid S_i), \dots, P(T_k \mid S_i)),$$

where $P(T_j|S_r)$ stands for the probability of topic T_j given sentence S_r.

Using the Bayes rule, we have

$$P(T_j \mid S_r) = \frac{P(S_r \mid T_j) * P(T_j)}{P(S_r)}$$

(1)

where

- $P(S_r|T_j)$ stands for the probability that topic T_j generates sentences S_r.
- $P(T_j)$ stands for the probability of Topic T_j.
- $P(S_r)$ stands for the probability of Sentence S_r.

Let's assume that a sentence S_r of a document D_b represents a topic T_j if the topic T_j generates all the words of the sentence S_r with some probability and that the document D_b generates Topic T_j. Under this assumption, we have:

$$P(S_r \mid T_j) = \sum_{W_i \in S_r} P(W_i \mid T_j) * P(D_b \mid T_j) * P(D_b) \qquad (2)$$

where

- $P(W_i|T_j)$ stands for the probability that topic T_j generates word W_i.
- $P(D_b|T_j)$ stands for the probability that document D_b generates Topic T_j.
- $P(D_b)$ stands for the probability of document D_b.

Given both $P(T_j|D_b)$ and $P(W_i|T_j)$ as the output of the LDA model, we can calculate the probability $P(D_b|T_j)$ (the probability of being topic T_j given document D_b) using the Bayes rule:

$$P(D_b \mid T_j) = \frac{P(T_j \mid D_b) * P(D_b)}{P(T_j)} \qquad (3)$$

Moreover, we can calculate the probability $P(S_r|T_j)$ by applying Equation 3 to Equation 2:

$$P(S_r \mid T_j) = \frac{P^2(D_b)}{P(T_j)} * \sum_{W_i \in S_r} P(W_i \mid T_j) * P(T_j \mid D_b) \qquad (4)$$

Furthermore, by first applying Equation 4 to Equation 1 and then normalizing the probability according to the length of sentence S_r, we can get:

$$P(T_j \mid S_r) = L * \frac{\sum_{W_i \in S_r} P(W_i \mid T_j) * P(T_j \mid D_b)}{Length\ (S_r)} \qquad (5)$$

where $L = \frac{P^2(D_b)}{P(S_r)}$ is a constant since both $P(D_b)$ and $P(S_r)$ are constant. In our case, for the sake of simplicity, we set L=1.

Finally, let's denote the summary's topic probability distribution as $P(T \mid summ)$. If there are m sentences in the current summary, i.e. ($S_1, S_2, \dots, S_m \in summ$), each component of the probability distribution can be averaged over all the m sentences in the summary:

$$P(T_j \mid summ) = \frac{1}{m} \sum_{i=1}^{m} P(T_j \mid S_i)$$

B. Summary generation

Here, we explore two methods in generating the summary: the static method and the dynamic method. While the former is employed to detect the salience of information in a static way, the later further controls redundancy in a dynamic way.

Static method

The similarity is calculated between each sentence and the given multi-documents via the topic probability distribution. In addition, the sentences ranked highest are extracted to produce the summary in the following way:

[1] http://sourceforge.net/projects/jgibblda

1) Run LDA for a fixed number of topics k. Here we use two ways to apply LDA, **SingleModel** and **MultiModel**, as described in Section III.
2) Get $P(T_j|D_k)$ and $P(W_i|T_j)$ from the trained LDA models and use them to calculate the similarity between a sentence and the given multi- documents.
3) Order the sentences according to the similarity score.
4) Pick up the sentences in descending order and include them in the summary until the summary has reached the size limitation.

Dynamic method

We propose a dynamic method to remove redundancy, which calculates the similarity between the summary and the given multi-documents in a dynamic way, and augments the summary incrementally. In particular, the sentence which makes the summary most similar to the given multi-documents is picked up to augment the summary. Compared with the above static method, the dynamic method generates the summary in the following dynamic way:

1) Run LDA, get $P(T_j|D_k)$ and $P(W_i|T_j)$ and order the sentences according to the similarity score, in the same way as the static method.
2) Include the top-scoring sentence as the initial summary.
3) Pick the sentence which maximizes the similarity score between the augmented summary and the given multi-documents.
4) Repeat last step until the summary has reached the size limitation.

C. Summary compression

In order to improve the coverage of the summary, summary or sentence compression is necessary for extractive summarization. In this paper, we compress a sentence by keeping the semantic arguments of main verbal predicate in the sentence via a shallow semantic parser. In particular, a state-of-the-art semantic role labeling (SRL) toolkit (Li et al. 2009) is employed to label various kinds of semantic arguments, such as agent, target, instrument, manner, location, time, etc. for a predicate. Here, we only consider the main verbal predicate which occurs at the highest level of syntactic parse tree. For simplicity, we perform sentence compression in the pre-processing stage for all sentences in the corpus:

1) Parse each sentence in the corpus using Berkeley parser.
2) Label semantic arguments of the main verbal predicate for each parsed sentence.
3) For each sentence, retain only those constituents which act as semantic arguments of the main verbal predicate, and delete all the remaining constituents.

D. Employing additional features

One main advantage of Our MDS framework lies in its simplicity. In this paper, we explore various popular features widely used in the literature, including the sentence position, the sentence length, the cosine similarity between the sentence and the document title, and the sentence's TF-IDF value.

Finally, we score a sentence by linearly interpolating the above feature scores with the original KL score as follows:

$$\begin{aligned} Score\ (S_r) = {} & w_0 KL\ (S_r, Doc\ ) + w_1 Pos\ (S_r) \\ & + w_2 Len\ (S_r) + w_3 CosSim\ \ (S_r, Tit\ ) \\ & + w_4 Tfidf\ \ (S_r) \end{aligned}$$

where w_0, w_1, w_2, w_3 and w_4 are fine-tuned to 1.1, 2, 0.3, 0.9, 0.3 respectively.

V. EXPERIMENTATION

We have systematically evaluated our MDS framework on the update summarization task of the Text Analysis Conference (TAC) 2008. It is worth to mention here that we only consider generic summarization and ignore the update nature of this task.

A. Experimental Setting

The documents for the TAC 2008 update summarization task are retrieved from the AQUAINT-2 collection of newswire articles. There are 48 datasets, each of which contains 20 documents. Besides, the retrieved documents are ordered chronologically and divided into two sets of 10 documents each. The task is to produce a 100-word summary for each set, guided by a statement describing the reader's need for information.

For evaluation, we use the ROUGE toolkit (Lin and Hovy 2003) provided by the TAC 2008 update summarization task[2]. Rouge-1, Rouge-2 and Rouge-SU4 scores at the 95% confidence level are computed by running ROUGE-1.5.5 with stemming but without the removal of stop words, where Rouge-2 and Rouge-SU4 are automatic ROUGE evaluation scores in TAC 2008. In order to compare our method with others, we extract a 100-word summary such as required by TAC 2008.

B. Experimental Results

Table I presents various Rouge scores using two LDA models trained in different ways and using different summarization methods on Set A and Set B. Each cell contains the results for both Set A and B (separated by "/"), as each cell in the following tables does. It shows that MultiModel performs better than SingleModel. This indicates that the obtained topic probability distribution is more accurate when LDA is applied on a small-scale relevant data. When applying LDA on each dataset, the derived topic probability distribution may better reflect the natural distribution of all the topics. In addition, for the specific dataset, Table I shows that the dynamic method outperforms the static method as expected due to redundancy control.

Since our framework performs better when applying LDA on each dataset, we use MultiModel by default in the following experiments. Table II show the effect of summary compression and employing popular features using the static and dynamic methods. The results show that both summary compression and employing popular features improve the performance. Unsurprisingly, the popular features greatly improve the performance. This indicates that these features are useful and can be well integrated into our framework. To our disappointment, summary compression via SRL only slightly improves the performance due to only consideration on long sentences and imperfect semantic role labeling. Moreover, Table III

[2] http://www.nist.gov/tac/data/index.html

presents the Rouge scores by using the LDA model with the removal of stop words. Compared with the experimental results in Table II, Table III shows that the removal of stop words achieves better Rouge scores. This indicates that stop words may bring some noise into the LDA model and affect the system performance.

In order to evaluate the performance of redundancy removal, we also include a statistical baseline to avoid repetition of words in the summary, motivated by the work of Larkey et al. (2003). Here, we use two features: one is the number of words repeated and another is the cosine similarity of TF-IDF between two sentences. Table IV shows the performance of the statistical method in removing the redundancy of the summary. It shows that statistical redundancy removal slightly improves the performance. Table IV also shows that the dynamic method outperforms the statistical method on Set A and performs comparable on Set B (much due to the update nature of the Set B). Finally, it shows that the statistical method fails to complement the dynamic method.

Compared to the official experimental results (averaged over Set A and Set B) published by TAC organizers (Dang and Owczarzak, 2008), our topic-driven MDS framework shows promising results: our best Rouge-2 score (0.0982) ranks No.3 and the best Rouge-SU4 score (0.1336) ranks No.5. However, the performance gap is quite small. Compared to the best system on the TAC 2008 update summarization task (Gillick et al. 2008) with the Rouge-2 score of 0.1038 and Rouge-SU4 score of 0.1362), our system only performs slightly low.

TABLE I. MDS PERFORMANCE ON SET A/SET B USING TWO LDA MODELS

	Rouge-1	Rouge-2	Rouge-SU4
SingleModel			
Static	0.3371/0.3356	0.0707/0.0733	0.1139/0.1163
Dynamic	0.3467/0.3428	0.0727/0.0735	0.1221/0.1230
MultiModel			
Static	0.3512/0.3507	0.0804/0.0789	0.1212/0.1214
Dynamic	0.3540/0.3533	0.0831/0.0853	0.1288/0.1294

TABLE II. CONTRIBUTION OF ADDITIONAL FEATURES AND SUMMARY COMPRESSION ON SET A/SET B

	Rouge-1	Rouge-2	Rouge-SU4
Static	0.3512/0.3507	0.0804/0.0789	0.1212/0.1214
+compression	0.3513/0.3503	0.0815/0.0803	0.1223/0.1226
+Features	0.3552/0.3543	0.0906/0.0895	0.1264/0.1272
+Features +compression	0.3556/0.3531	0.0932/0.0952	0.1286/0.1287
Dynamic	0.3540/0.3533	0.0831/0.0853	0.1288/0.1294
+compression	0.3542/0.3534	0.0881/0.0888	0.1291/0.1299
+Features	0.3615/0.3611	0.0898/0.0914	0.1301/0.1315
+Features +compression	0.3618/0.3615	0.0955/0.0960	0.1305/0.1316

TABLE III. MDS PERFORMANCE ON SET A/SET B USING LDA MODEL WITH THE REMOVAL OF STOP WORDS

	Rouge-1	Rouge-2	Rouge-SU4
Static	0.3623/0.3606	0.0926/0.0997	0.1286/0.1331
Dynamic	0.3699/0.3609	0.0961/0.0996	0.1332/0.1336

TABLE IV. PERFORMANCE OF REDUNDANCY REMOVAL

	Rouge-1	Rouge-2	Rouge-SU4
Statistical	0.3628/0.3618	0.0930/0.1002	0.1288/0.1339
Dynamic	0.3699/0.3609	0.0961/0.0996	0.1332/0.1336
Dynamic + Statistics	0.3699/0.3613	0.0960/0.0998	0.1331/0.1337

VI. CONCLUSION

In this study, we present a simple topic-driven framework in extracting a generic summary from multi-documents based on the probability distribution over the derived topics, which applies to various kinds of text units, such as word, sentence, summary, document and multi-documents. The intuition behind our approach is that the summary should have similar topic probability distribution with that of given multi-documents. Evaluation on the TAC 2008 corpus shows promising results. In addition, further enhancements are used to improve the quality of summarization, such as employing popular features and summary compression. This indicates the flexibility of our topic-driven framework. In the future, we will adopt our framework on DUC-style summarization.

ACKNOWLEDGMENT

This research was supported by Projects 60970056 under the National Natural Science Foundation of China.

REFERENCES

[1] Arora, R. and Ravindran B. 2008. Latent dirichlet allocation and singular value decomposition based multi-document summarization. *Proceedings of ICDM 2008.* pages 713-718.

[2] Bhandari, H., Shimbo M., Ito T., and Matsumoto Y. 2008. Generic text summarization using probabilistic latent semantic indexing. *Proceedings of IJCNLP 2008*, pages 133-140.

[3] Blei, D. M., Ng A. Y., and Jordan M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, pages 993-1022.

[4] Dang, H. T. and Owczarzak K. 2008. Overview of the TAC 2008 Update Summarization Task. *Proceedings of the First Text Analysis Conference (TAC 2008)*. Maryland, USA. Pages:1-21.

[5] Gildea, D. and Jurafsky D. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3): 245-288.

[6] Gillick, D., Favre B., and Hakkani-Tur D. 2008. The ICSI Summarization System at TAC 2008. *Proceedings of the First Text Analysis Conference (TAC 2008)*. Maryland, USA. Nov. 17-19.

[7] Haghighi, A. and Vanderwende L.2009. Exploring Content Models for Multi-Document Summarization. *The 2009 Annual Conference of the North American Chapter of the ACL*, pages: 362-370.

[8] Larkey, L.S., Allan J., Connell M.E., Bolivar A., and Wade C. 2003. UMass at TREC 2002: Cross Language and Novelty Tracks. *National Institute of Standard & Technology.* Pages: 721-732.

[9] Li, J.H., Zhou G.D., Zhao H., Zhu Q.M., and Qian P.D. 2009. Improving Nominal SRL in Chinese Language with Verbal SRL and Automatic Predicate Recognition. *Proceedings of the Empirical Methods for Natural Language Processing (EMNLP '09)*, Singapore, Aug. 6-7. Pages: 1280-1288.

[10] Lin, C.Y. and Hovy E. H. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton. Canada.

[11] Radev, D. R., Jing H. and Budzikowska M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *In ANLP- NAACL Workshop on Summarization*, Seattle, WA.

[12] Wang, D., Zhu S., Li T. and Gong Y. 2009. Multi-document summarization using sentence-based Topic Models. *Proceedings of the ACL-IJCNLP 2009 Conference Short Paper,* Pages 297-300.