Computer Science and Information Systems 12(4):1375–1389 DOI: 10.2298/CSIS140815060N

# Topic-Sensitive Multi-document Summarization Algorithm

Liu Na[1], Di Tang[2], Lu Ying[1], Tang Xiao-jun[1] and Wang Hai-wen[1]

[1] School of Information Science & Engineering, Dalian Polytechnic University, 116034 Dalian, China
liuna@dlpu.edu.cn
[2] School of Computer and Information Technology, Liaoning Normal University, 116029 Dalian, China

**Abstract.** Latent Dirichlet Allocation (LDA) has been used to generate text corpora topics recently. However, not all the estimated topics are of equal importance or correspond to genuine themes of the domain. Some of the topics can be a collection of irrelevant words or represent insignificant themes. This paper proposed a topic-sensitive algorithm for multi-document summarization. This algorithm uses LDA model and weight linear combination strategy to identify significance topic which is used in sentence weight calculation. Each topic is measured by three different LDA criteria. Significance topic is evaluated by using weight linear combination to combine the multi-criteria. In addition to topic features, the proposed approach also considered some statistics features, such as term frequency, sentence position, sentence length, etc. It not only highlights the advantages of statistics features, but also cooperates with topic model. The experiments showed that the proposed algorithm achieves better performance than the other state-of-the-art algorithms on DUC2002 corpus.

**Keywords:** multi-document summarization, LDA, topic model, weighted linear combination.

## 1. Introduction

Over the past several years, there has been much interest in the task of multi-document summarization. The main task of multi-document summarization is to extract the most important sentences from multiple documents and format them into a summary. Therefore, finding an appropriate method to justify the importance or relevance of sentence dominates this research area. Many proposed approaches use statistical methods, lexical chains, graph-based algorithms, or Bayesian language models to produce summaries.

Recently, generative models for documents have begun to explore topic-based content representations approaches. In natural language processing, Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. If observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one

of the document's topics. LDA topic model can back to the early 2002s. Since then, topic model has been discovered, re-discovered, and extended many times in different communities. For example, Michal Rosen-Zvi introduced an author-topic model in 2004. The model was a generative model for documents that extends LDA to include authorship information [1]. Aria Haghighi utilized a hierarchical LDA-style model in 2009. The documents set content was represented specificity as a hierarchy of topic vocabulary distributions [2]. Other methods based on topic mode for natural language processing have also been presented. He Tingting proposed a multi-aspect Blog sentiment analysis method using LDA topic model and Hownet lexicon in 2012 [3]. Jean-Yves Delort intruduced an unsupervised probabilistic approach based on topic-model, called Dualsum, for update summarization in 2012 [4]. Bassam Al-Salemi presented a boosting algorithm called AdaBoost.MH for multi-label classification in 2015 [5]. Ximing Li illustrated an extension of L-LDA, namely supervised labeled latent Dirichlet allocation (SL-LDA), for document categorization in 2015 [6].

Most of these methods use LDA to model each document as a mixture of probabilistic topics. However, the research on the number of topic is limited. In fact, the setting of topic number can affect the interpretability of the results. Models with very few topics would result in broad topic dentitions that could be a mixture of two or more distributions. On the other hand, models with too many topics are expected to have very specific descriptions that are uninterruptible [7]. Generally the more similar the sentence with document, the more likely the sentence is selected into summarization. It is essential to identify which topic is meaningful to the thematic structure of document before sentence selection before sentence selection.

For generic multi-document summarization, we propose a topic-sensitive multi-document summarization algorithm. The proposed algorithm not only uses topic features of sentences, but also utilizes statistical features of sentences. First, it introduces and defines the concept of significance topic. Second, it separates topic into significance and insignificance by the multi-criteria measures. Third, it calculates sentence score with significance topic and statistics features. And finally, it constructs multi-document summarization according sentence score ranking.

The paper is organized as follows. Section 2 introduces related work. Section 3 describes the Latent Dirichlet allocation algorithm. Section 4 defines the concept of significance topic. Section 5 discusses generative Topic-sensitive multi-document summarization algorithm, and Section 6 presents the results of applying this algorithm to DUC datasets. We conclude and discuss further research directions in Section 7.

## 2. Related Work

In this section, we firstly introduce some representative work about multi-document summarization. Then, we present some literatures which relevant to topic-based multi-document summarization algorithm.

The research of automatic text summarization has continued more than 50 years. Many approaches have been addressed and many solutions have been evaluated. The most popular extractive summarization methods always score sentences based on features such as word frequency, TF-IDF weighting, sentence position, title relation and cue-phrases. All these approaches are built based on the features of documents without

considering the semantic associations behind sense. Other approaches take account of semantic associations between words and combine them with those features in the process of sentence similarity. Examples of such approaches are: latent semantic analysis [8], topic signatures [9], sentence clustering [10], and Bayesian topic model based approaches, such as BayeSum [11], topic segmentation [12], and TopicSum from [13], and so on. Although these approaches can enhance performance of retrieval and document summarization significantly, these approaches ignore contextual information of words, which can significantly influence overall performance of sentence similarity.

Especially, we are mainly inspired by following pioneering work. Recently, many approach for multi-document summarization based on topic model has been presented. Dingding Wang presented a new Bayesian sentence-based topic model for summarization in 2009. This model made use of both the term-document and term-sentence associations to help the context understanding and guide the sentence selection in the summarization procedure [14]. Liu S presented an enhanced topic modeling technique in 2012. This technique provided users a time-sensitive and more meaningful text summary [15]. WY Yulong proposed SentTopic-MultiRank, a novel ranking model for multi-document summarization in 2012. This method assumed various topics to be heterogeneous relations, and then treated sentence connections in multiple topics as a heterogeneous network, where sentences and topics were effectively linked together [16]. Li Jiwei proposed a novel supervised approach taking advantages of both topic model and supervised learning in 2013. This approach incorporated rich sentence feature into Bayesian topic models [17]. Sanghoon Lee proposed a new multi-document summarization method that combines topic model and fuzzy logic model in 2013. The method extracted some relevant topic words by topic model and uses them as elements of fuzzy sets. The final summarization was generated by a fuzzy inference system [18]. Zhang R introduced a novel speech act-guided summarization approach in 2013. This method used high-ranking words and phrases as well as topic information for major speech acts to generate template-based summaries [19]. Zhu Y presented a novel relational learning-to-rank approach for topic-focused multi-document summarization in 2013. This approach incorporated relationships into traditional learning-to-rank in an elegant way [20]. Tan Wentang introduced a generative topic model PCCLDA(partial comparative cross collections LDA) for multi-collections in 2013. This approach detected both common topics and collection-special topics, and modeled text more exactly based on hierarchical dirichlet processes [21]. Bian J introduced a new method of sentence-ranking in 2014. The method combined topic-distribution of each sentence with topic-importance of the corpus together to calculate the posterior probability of the sentence, and then, based on the posterior probability, it selected sentences to form a summary [22]. Zhou S proposed an automatic summarization algorithm based on topic distribution and words distribution in 2014. The algorithm was a fully sparse topic model to solve the problem of sparse topics in muti-document summarization [23]. Guangbing Yanga proposed a novel approach based on recent hierarchical Bayesian topic models in 2015. The proposed model incorporated the concepts of n-grams into hierarchically latent topics to capture the word dependencies that appear in the local context of a word. The quantitative and qualitative evaluation results showed that this model has outperformed both hLDA and LDA in document modeling [24].

The success of these models and applications suggest that the mechanism of incorporating the concept of latent topics into n-grams is helpful for the problems of multi-document summarization. Indeed, a similarity between these literatures with our

current idea lies in that we all consider using LDA topic model to present documents. However, we do not aim to use the overall topic distributed from documents simply. In their work, the research on the number of topic is limited. In fact, the setting of topic number can affect the interpretability of the results. Differently, our method identifies significance topic and use significance topic to calculate sentence score. Further, we make full use of traditional feature of sentences to improve the overall sentence ranking performance.

## 3. Latent Dirichlet Allocations

In this section we describe LDA. LDA is a generative probabilistic model for a corpus. The basic idea of LDA is that documents are represented as random mixtures over latent topics, each of which is characterized by a distribution over words.

The LDA model is represented as a probabilistic graphical model in Fig. 1. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. As depicted in the figure, there are three levels to the LDA representation. The parameters $\alpha$ and $\beta$ are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables $\theta_d$ are document-level variables, sampled once per document. Finally, the variables $z_{dn}$ and $w_{dn}$ are word-level variables and are sampled once for each word in each document.

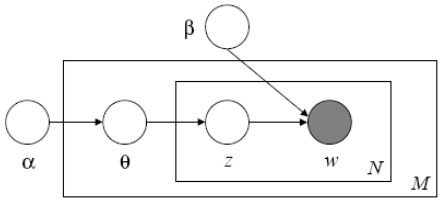


**Fig. 1.** Graphical model representation of LDA

LDA assumes the following generative process for each document $w$ in a corpus $D$:

(1) Choose $N$ ～Poisson ($\xi$).
(2) Choose $\theta$ ～Dir ($\alpha$).
(3) For each of the $N$ words $w_n$:
  (a) Choose a topic $z_n$ ～Multinomial ($\theta$).
  (b) Choose a word $w_n$ from $p$ ($w_n|z_n,\beta$), a multinomial probability conditioned on the topic $z_n$.

A $k$-dimensional Dirichlet random variable $\theta$ can take values in the ($k$-1)-simplex, and has the following probability density on this simplex:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Pi_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{1}$$

where the parameter $\alpha$ is a $k$-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient distribution on the simplex, has finite

dimensional sufficient statistics, and is conjugate to the multinomial distribution. Given the parameters $\alpha$ and $\beta$, the joint distribution of a topic mixture $\theta$, a set of $N$ topics $z$, and a set of $N$ words $w$ is given by:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{2}$$

where $p(z_n \mid \theta)$ is simply $\theta_i$ for the unique $i$. Integrating over $\theta$ and summing over $z$, we obtain the marginal distribution of a document:

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) (\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta)) d\theta \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) (\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta)) d\theta_d \tag{4}$$

In LDA, the observed data are the words of each document and the hidden variables represent the latent topical structure. In a given dataset, the posterior distribution can determines a hidden topical decomposition of the dataset. Many topic model applications use posterior estimates to perform tasks, such as information retrieval and document browsing.

## 4. Significance Topic of LDA Model

LDA is a statistical generative model that represents documents as a mixture of probabilistic topics and topics as a mixture of words. However, the inferred topics of LDA do not always represent meaningful themes. The setting of the number of topic $K$ is extremely critical. It can affect the quality of the model directly. Therefore, it is essential to identify significance and insignificance topics before calculation the similarity between sentence and document.

To identify significance topics from LDA model, the following learning setting is considered. Given a dataset of $D=\{D_1,D_2,\ldots,D_M\}$ documents with a total of $N$ token words and $W=\{W_1,W_2,\ldots,W_N\}$ unique terms and $R$ sentence $S=\{S_1,S_2,\ldots,S_R\}$, a topic model $T=\{T_1,T_2,\ldots,T_K\}$ is generated from fitting its parameters, $\theta$ and $\varphi$, to the dataset assuming that the number of topics is set to $K$.

### 4.1. Significance Topic Definition

In this section, we introduce the distance between topic and word, topic, document. These three distances is used to evaluate the significance of a topic.

***Definition 1.*** Diversity of word distribution

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. For example, the probability of vocabulary "casualties" and "victims" could be far greater than other words in the document describing the topic "earthquake". Therefore, if the probability *P(W|T)* of word *W* under topic *T* is approximate average, then the diversity of word distribution is lower. We estimate the diversity by KL (Kullback–Leibler) Divergence.

Assumption the probability distribution *P(T|D)* of topics for each document and probability distribution *P(W|T)*of the words over the topic, Diversity of topic-word distribution for $W_i$ over topic $T_k$ is defined as

$$D_k^{\alpha} = KL(P(W_i \mid T_k) \| P(W_i \mid T_s)) = \sum_t P(W_i \mid T_k) \log_2 \frac{P(W_i \mid T_k)}{P(W_i \mid T_t)} \tag{5}$$

***Definition 2.*** Diversity of topic distribution

Each document is always a subset of collection which is relevant to the same topic in multiple document collections. If the probability *P(T|D)* of topic *T* under document *D* is approximately equal, then the diversity of topic distribution is lower. We estimate the diversity by KL Divergence

$$D_k^{\beta} = KL(P(T_k \mid D_p) \| P(T_k \mid D_M)) = \sum_d P(T_k \mid D_p) \log_2 \frac{P(T_j \mid D_p)}{P(T_j \mid D_d)} \tag{6}$$

***Definition 3.***Similarity of Topic distribution

If the distribution of topic *k* is completely different with all other topics, it must not be considered as most talked topics. But KL divergence is not symmetric, thus the similarity of topic distribution is used Jensen-shannon distance [25]:

$$\begin{aligned} D_k^{\gamma} &= \frac{1}{2}[KL(P(T_i \mid T_k) \| P(T_i \mid T_s)) + KL(P(T_k \mid T_i) \| P(T_i \mid T_s))] \\ &= \frac{1}{2}[\sum_t P(W_i \mid T_k) \log_2 \frac{P(W_i \mid T_k)}{P(W_i \mid T_t)} + \sum_t P(W_k \mid T_i) \log_2 \frac{P(W_k \mid T_i)}{P(W_i \mid T_t)}] \end{aligned} \tag{7}$$

### 4.2. Multi-criteria Measures for Significance Topic

In this paper, for each topic *k*, three different categories of topic significance criteria $C = (\alpha, \beta, \gamma)$ are determined. In order to combine the multi-criteria measures, we use Weighted Linear Combination (WLC) decision strategy to form a single index of evaluation. The simplest form of WLC is as follows:

$$A_k = \psi_k S_k \tag{8}$$

where $\psi_k$ is weight of total score of topic *k*, $S_k$ is score of topic *k* under criteria *C*.

To construct the score and weight for each topic, we transfer each true value into two standardized scores, one is a relative score $S_k$ and the other is a weight value $\psi_k$ between 0 and 1.

(1)Normalize $C = (\alpha, \beta, \gamma)$ for each topic

$D1_k^C$ denotes the topic scores based on the weight of each score with respect to the total score over all topics.

$$D1_k^C = D_k^C \times \frac{\sum_{j=1, j\neq k}^{K} D_j^C}{\sum_{j=1}^{K} D_j^C} \tag{9}$$

$D2_k^C$ denotes topic weights between 0 and 1 using minimum and maximum values.

$$D2_k^C = \frac{D_k^C - D_{\min}^C}{D_{\max}^C - D_{\min}^C} \tag{10}$$

(2) Combine the three criteria using WLC

In this paper, we assume that three criterion are equally. Thus, the score $S_k$ in the Equation (8) is defined as mean score of three normalized score.

$$S_k = \frac{1}{3} D1_k^C \tag{11}$$

The weights $\psi_k$ in the Equation (8) determine the contributions in the total score.

$$\psi_k = \lambda_\alpha D2_k^\alpha + \lambda_\beta D2_k^\beta + \lambda_\gamma D2_k^\gamma \tag{12}$$

where $\lambda_\alpha, \lambda_\beta, \lambda_\gamma$ is weight of criteria *C*. In our experiments, $\lambda_\alpha = 0.25, \lambda_\beta = 0.25, \lambda_\gamma = 0.5$.

(3) Compute the final score $TS_k$ for each topic *k*

From $TS_k$, we can estimate the significance topic.

$$TS_k = S_k \times \psi_k \tag{13}$$

## 5. Topic-Sensitive Multi-document Summarization Algorithm

Most multi-document summarization algorithm adopted statistical features of sentence, such as word frequency, position information, as a basis for sentence weighting, while seldom considering lexical semantic relation. In this paper, in additional to these traditional features, our approach also uses LDA feature to calculate sentence weight.

### 5.1. Statistical Features of Sentence

Traditional sentence features include term frequency, sentence position and sentence length, etc.

(1) Word Frequency

The word frequency is often used for calculating the importance of sentence. The more feature word the sentence contained, the more information the sentence contained. The word frequency score $F(S_{d,i})$ of sentence $S_{d,i}$ in document $D_i$ is measured as

$$F(S_{d,i}) = \frac{TF * IDF(S_{d,i})}{TF * IDF_{avg}(D_i)} \tag{14}$$

where $TF*IDF(S_{d,i})$ is sum of term frequency–inverse document frequency of word of sentence $S_{d,i}$ in document $D_i$, $TF*IDF_{avg}(D_i)$ is average of $TF*IDF(S_{d,i})$ of document $D_i$.

(2) Sentence Length

In summarization, too long or too short sentences are not expected. The sentence length score $L(S_{d,i})$ of sentence $S_{d,i}$ in document $D_i$ is measured as

$$L(S_{d,i}) = 1 - \frac{\left| l(S_{d,i}) - l_{avg}(D_i) \right|}{l_{avg}(D_i)} \tag{15}$$

where $l(S_{d,i})$ the number of word of sentence $S_{d,i}$, $l_{avg}(D_{,i})$ is the average length of document $D_i$.

(3) Sentence Position

Sentence position always gives the importance of the sentences. Research showed that the proportion that the first sentence as summarization is 85%, and the last sentence as summarization is 7%. The sentence position score $P(S_{d,i})$ of sentence $S_{d,i}$ in document $D_i$ is measured as

$$P(S_{d,i}) = \begin{cases} 0.8 & (S_{d,i} \ \ is \ \ the \ \ first \ \ sentence) \\ 0.2 & (S_{d,i} \ \ is \ \ the \ \ last \ \ sentence) \\ 0 & other \end{cases} \tag{16}$$

According to the above three features, the sentence weight is calculated as

$$SCORE\_statistics(S_{d,i}) = F(S_{d,i}) + L(S_{d,i}) + P(S_{d,i}) \tag{17}$$

## 5.2. LDA Features of Sentence

In general, similarity between sentences is measured by the number of co-occur word. But it ignores the semantic relevance. This section introduces how to calculate similarity between sentences with document based on topic model.

(1) Sentence Topic

We break down the set of documents into topics by LDA model. We assume that a sentence $S_r$ in document $D_m$ represents a topic $T_j$ if all the words of the sentence $S_r$ belong to the topic $T_j$ and that the topic belongs to the document $D_m$. Thus topic of sentence is measured by sum of word probability.

$$P(S_r \mid T_j) = \prod_{W_i \in S_r} P(W_i \mid T_j) * P(T_j \mid D_m) * P(D_m) \tag{18}$$

where $P(S_r \mid T_j)$ stands for probability that a sentence $S_r$ represents topic $T_j$, $\prod_{W_i \in S_r} P(W_i \mid T_j)$ stands for probability that the words of $S_r$ belongs to topic $T_j$, $P(T_j \mid D_m)$ stands for probability that topic $T_j$ belongs to document $D_m$, $P(D_m)$ stands for probability of document $D_m$.

In this paper, for the sake of simplicity, we have assumed that all documents are equi-probable. Thus the probability of the document values don't make any difference during the calculation of the $P(S_r \mid T_j)$. Thus $P(S_r \mid T_j)$ is modified as

$$P(S_r \mid T_j) = \prod_{W_i \in S_r} P(W_i \mid T_j) * P(T_j \mid D_m) \tag{19}$$

Since $P(W_i \mid T_j) < 1$, $\prod_{W_i \in S_r} P(W_i \mid T_j)$ is unfit for long sentences. With the definition of LDA, $P(W_i \mid T_j)$ is the distribution of word over topic, we replace $\prod_{W_i \in S_r} P(W_i \mid T_j)$ as $\sum_{W_i \in S_r} P(W_i \mid T_j)$. Thus Equation (19) is adapted as

$$P(S_r \mid T_j) = \sum_{W_i \in S_r} P(W_i \mid T_j) * P(T_j \mid D_m) \tag{20}$$

In Equation (20), the longer the sentences are, the higher the $P(S_r \mid T_j)$ is. Thus we normalize Equation (20) by dividing the length of sentence $S_r$

$$P(S_r \mid T_j) = \frac{\sum_{W_i \in S_r} P(W_i \mid T_j) * P(T_j \mid D_m)}{length(S_r)} \tag{21}$$

(2) Topic Similarity between Sentence and Document

LDA model can achieve the document distributions over topics and the topic distributions over words by considering Dirichlet priors. From Section 3, we can identify significance topic. Therefore, similarity between sentence and document is computed by KL Divergence between sentence and significance topic

$$KL(S_r, D_n) = \sum P(S_r \mid T_j) \log \frac{P(S_r \mid T_j)}{P(T_j \mid D_n)} \tag{22}$$

The LDA features of sentence is defined by $SCORE\_LDA(S_{d,i}) = KL(S_r, D_n)$.

(3) Algorithm

Now we would like to state the topic-sensitive multi-document summarization algorithm. The algorithm includes three main steps. First, pick significance topic according LDA model. Second, calculate sentences weight cooperating traditional features and LDA features. And finally, form summarization based on sentence ranking. The algorithm is described in detail as follow:

(a). Run the LDA model to get the probability distributions $P(T_j|D_k)$-probability of topic $T_j$ given document $D_k$ and $P(W_i|T_j)$-probability of word $W_i$ given topic $T_j$;

(b). Calculate $TS_k$ using probability distribution and Equation (13), pick significance topic;

(c). Calculate statistical features weight $SCORE\_statistics(S_{d,i})$ of each sentence according Equation(17);

(d). Calculate LDA features weight $SCORE\_LDA(S_{d,i})$ of each sentence according Equation(22), the final sentence weight is

$$SCORE(S_{d,i}) = SCORE\_statistics(S_{d,i}) + SCORE\_LDA(S_{d,i})$$

(e). Form summarization according sentence weight ranking.

## 6. Experimental Results

We use DUC2002 Corpus datasets (http://duc.nist.gov/) in our experiments. The data is made up of 59 common sets of documents in English. Each set has between 5 and 15 documents, with an average of 10 documents which are produced using data from the TREC disks used in the question-answering track in TREC-9. For the multi-document summarization tasks, DUC2002 corpus provides 200- and 400-word extracts as reference. To compare our experimental results with reference of DUC2002, we produce a generic summarization of documents with a length of approximately 200- and 400-words or less length.

We compare the proposed algorithm with term frequency algorithm (SumBasic algorithm [26]) and two other LDA algorithms (Doc-LDA [27] and KL-LDA [28]). The reason why we choose three methods as baseline lies in: baseline SumBasic algorithm is based on word frequency method to generate multi-document summarization, Doc-LDA and KL-LDA method are based on the topic model to generate multi-document summarization. Compared with SumBasic algorithm, we can verify whether topic model can improve the result of multi-document summarization. Compared with Doc-LDA and KL-LDA method, we can validate whether our proposed method is more effective than other method which is also based on LDA.

We perform our summarization results with keeping stop-words and removing stop-words. In both the cases Porter Stemmer is used to stem the words to their root form. We calculate the ROUGE scores separately for 200 and 400 length summary. Each experimental result is an average over 10 trials. The number of word-topics is equal to 50.

We are mainly interested in the ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-S4 and ROUGE-SU4 recall score. ROUGE is a widely used and standard measure for comparing the performance of summarization. The measure evaluates the quality of the summarization by counting the number of overlapping units, such as n-grams and word sequences, between the model summarization and candidate summarization.

Table 1 presents the results of keeping stop-words versus removing stop-words in terms of 200 words summarization. Table 2 presents the results of 400 words summarization.

A number of conclusions can be drawn from the results.

(1) Above algorithms based on LDA perform better than the term frequency algorithm. This holds for the summarization of length 200 and 400 words. It shows that LDA model can improve the quality of multi-document summarization.

(2) For the same algorithm, summarization results with stop words outperforms without stop words. The reason is that ROUGE evaluates the co-occurrence words units

between system summarization and expert summarization. Obviously, the increasing number of stop words leads to the increase of ROUGE score.

(3) The proposed algorithm gives better results in 200- and 400-words summarization. It shows that the proposed algorithm works irrelevant to the size of the summarization. Although our approach is also used the topic similarity between sentence and document, difference from other method, our approach identifies the significance topic and compute topic similarity between sentence and significance topic, which can avoid the adverse influence of insignificance topic. This is the reason why our LDA approach is outperformed the other two LDA approach.

**Table 1.** ROUGE score of 200 words summarization

| | algorithm | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-S4 | ROUGE-SU4 |
|---|---|---|---|---|---|---|
| keeping stop-words | SumBasic | 0.409 | 0.098 | 0.375 | 0.097 | 0.149 |
| | Doc-LDA | 0.434 | 0.153 | 0.391 | 0.146 | 0.194 |
| | KL-LDA | 0.408 | 0.133 | 0.374 | 0.129 | 0.176 |
| | proposed | 0.479 | 0.195 | 0.427 | 0.160 | 0.235 |
| removing stop-words | SumBasic | 0.309 | 0.074 | 0.296 | 0.060 | 0.102 |
| | Doc-LDA | 0.322 | 0.133 | 0.300 | 0.109 | 0.146 |
| | KL-LDA | 0.289 | 0.107 | 0.274 | 0.091 | 0.124 |
| | proposed | 0.368 | 0.170 | 0.339 | 0.142 | 0.168 |

**Table 2.** ROUGE score of 400 words summarization

| | algorithm | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-S4 | ROUGE-SU4 |
|---|---|---|---|---|---|---|
| keeping stop-words | SumBasic | 0.492 | 0.156 | 0.464 | 0.152 | 0.209 |
| | Doc-LDA | 0.535 | 0.254 | 0.504 | 0.246 | 0.294 |
| | KL-LDA | 0.470 | 0.190 | 0.448 | 0.188 | 0.236 |
| | proposed | 0.562 | 0.233 | 0.525 | 0.274 | 0.329 |
| removing stop-words | SumBasic | 0.378 | 0.119 | 0.364 | 0.095 | 0.143 |
| | Doc-LDA | 0.427 | 0.227 | 0.409 | 0.197 | 0.236 |
| | KL-LDA | 0.333 | 0.161 | 0.324 | 0.137 | 0.170 |
| | proposed | 0.453 | 0.257 | 0.436 | 0.229 | 0.278 |

For above five ROUGE scores, we focus on ROUGE-1 score. ROUGE-1 calculates the recall of the number of test summarization overlaps the model summarization with 1-gram word sequence. In experimental results, we compare the proposed algorithm with another four different algorithms in terms of ROUGE-1 recall measures and 95% Confidence Interval. Two of them are the top two algorithms of the DUC2002 Multi-Document Summarization task, ”Generating Single and Multi-Document Summaries with GISTEXTER” (GISTEXTER) [29] and ”Writing Style Recognition and Sentence Extraction” (WSRSE) [30]. Another two similar algorithms is based on LDA model. One only uses LDA to form summarization. Another uses LDA and SVD to form summarization [31]. The experimental results are shown in Table 3 and Table 4.

We also evaluate the proposed algorithm by considering both cases of keeping stop-words and removing stop-words for 200- and 400-words summarization. From Table 3 and Table 4, we can clearly see that the proposed algorithm performs better than

baseline. This can be attributed to the features that the algorithm presented in this paper uses the significance topic to improve the precision of topic model during process of constructing multi-document summarization.

**Table 3.** Recall for ROUGE-1 and 95% Confidence Interval of 200-words summarization

| | | GISTEXTER | WSRSE | LDA | LDA-SVD | Proposed algorithm |
|---|---|---|---|---|---|---|
| keeping stop-words | Recall | 0.487 | 0.487 | 0.556 | 0.561 | 0.570 |
| | 95% interval | 0.462-0.512 | 0.460-0.513 | 0.541-0.573 | 0.546-0.577 | 0.553-0.585 |
| removing stop-words | Recall | 0.395 | 0.401 | 0.456 | 0.459 | 0.465 |
| | 95% interval | 0.364-0.427 | 0.372-0.429 | 0.435-0.477 | 0.439-0.479 | 0.447-0.487 |

**Table 4.** Recall for ROUGE-1 and 95% Confidence Interval of 400-words summarization

| | | GISTEXTER | WSRSE | LDA | LDA-SVD | Proposed algorithm |
|---|---|---|---|---|---|---|
| keeping stop-words | Recall | 0.563 | 0.580 | 0.608 | 0.620 | 0.631 |
| | 95% interval | 0.547-0.580 | 0.556-0.602 | 0.597-0.619 | 0.609-0.631 | 0.618-0.641 |
| removing stop-words | Recall | 0.467 | 0.485 | 0.502 | 0.512 | 0.516 |
| | 95% interval | 0.447-0.488 | 0.460-0.512 | 0.486-0.518 | 0.498-0.528 | 0.489-0.537 |

## 7. Conclusions

Multi-document summarization algorithm based on sentence extraction mainly includes choosing sentences from the documents using some weighting mechanism and combining them into a summarization. This paper introduces a novel algorithm which is sensitive to topic for multi-document summarization. The proposed algorithm separates estimated topic into significance topic and insignificance topic. In term of sentence weight, we use similarity between sentence topic and significance topic as LDA features of sentence. Meanwhile, we utilize traditional features such as term frequency, sentence position and sentence length. In the future, we will consider how to determine the number of topic by significance topic automatically.

**Acknowledgment.** This work was supported by National Natural Science Foundation of China (NO. 61402069, NO.61272369, NO.61175053, NO.61301185, NO.61370070), General project of Liaoning Provincial Department of Education (NO.L2015047), Science and Technology Foundation of Dalian City (NO.2013J21DW006).

## References


1. Rosen-Zvi, M., Griffiths, T., Steyvers, M., et al.: The Author-Topic Model for Authors and Documents. In Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Banff, Canada, pp: 487-494. (2004)
2. Haghighi, A., Vanderwende, L.: Exploring Content Models for Multi-Document Summarization. In Proceedings of The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics,Colorado,US, pp: 362-370. (2009)
3. Tingting, H., Fang, L.: Semantic Knowledge Acquisition from Blogs with Tag-Topic Model. China Communications, vol. 9, no. 3: 38-48. (2012)
4. Delort, J.-Y., Alfonseca, E.: DualSum: a topic-model based approach for update summarization. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, pp: 214-223. (2012)
5. Al-Salemi, B., Juzaiddin Ab Aziz, M., Azman Noah, S.: LDA-AdaBoost.MH: Accelerated AdaBoost.MH based on latent Dirichlet allocation for text categorization. Journal of Information Science, vol. 41, no. 1: 27-40. (2015)
6. Li, X., Ouyang, J., Zhou, X., Lu, Y., Liu, Y.: Supervised labeled latent Dirichlet allocation for document categorization. Applied Intelligence, vol. 42, no. 3: 581-593. (2015)
7. Steyvers, M., Griths, T. L.: Probabilistic Topic Models, Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum. (2005)
8. Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp: 19 – 25. (2001)
9. Lin, C.-Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In Proceedings of the international conference on computational linguistic, pp: 495 – 501. (2000)
10. He, R., Qin, B., Liu, T.: A novel approach to update summarization using evolutionary manifold-ranking and spectral clustering. Expert Systems with Applications, vol.39, no.3: 2375 – 2384. (2012)
11. Daumé III, H., Marcu, D.: Bayesian query-focused summarization. In Proceedings of the conference of the association for computational linguistics (ACL) and 44th annual meeting of the ACL ,Sydney, pp: 305 – 312. (2006)
12. Eisenstein, J., Barzilay, R.: Bayesian unsupervised topic segmentation. In Proceedings of the conference on empirical methods in natural language processing, Honolulu, Hawaii, pp: 334-343. (2008)
13. Haghighi, A.,Vanderwende, L.: Exploring content models for multidocument summarization. In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, Colorado, pp: 362 – 370. (2009)
14. Wang, D., Zhu, S., Li. T.: Multi-document summarization using sentence-based topic models. In Proceedings of the ACL-IJCNLP 2009 Conference, Association for Computational Linguistics, Suntec,Singapore, pp: 297-300. (2009)
15. Liu, S., Zhou, M. X., Pan, S.: TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. Intelligent Systems & Technology, vol.3, no.2: 67-83. (2012)
16. Yulong, W. Y., Fan, P., Huang, Z. L.: SentTopic-MultiRank: a novel ranking model for multi-document summarization. Proceedings of International Conference on Computational Linguistics (COLING 2012), Mumbai, pp: 2977–2992. (2012)
17. Jiwei, L., Sujian, L.: A Novel Feature-based Bayesian Model for Query Focused Multi-document Summarization, Transactions of the Association for Computational Linguistics, pp: 89 – 98. (2013)

18. Lee, S., Belkasim, S., Zhang, Y.: Multi-document text summarization using topic model and fuzzy logic. In Proceedings of the 9th international conference on Machine Learning and Data Mining in Pattern Recognition, Springer-Verlag Berlin, New York, pp: 159-168. (2013)
19. Zhang, R., Li, W., Gao, D., et al.: Automatic Twitter Topic Summarization With Speech Acts. Audio Speech & Language Processing, vol. 21, no.3: 649 - 658. (2013)
20. Zhu, Y., Lan, Y., Guo, J., et al.: A Novel Relational Learning-to-Rank Approach for Topic-Focused Multi-document Summarization. IEEE International Conference on Data Mining, pp: 927 – 936. (2013)
21. Wentang, T., Zhenwen, W., Fengjing, Y., et al.: A Partial Comparative Cross Collections LDA Model. Computer Research and Development, vol.50, no. 9: 1943-1953. (2013)
22. Bian, J., Jiang, Z., Chen, Q.: Research on Multi-document Summarization Based on LDA Topic Model. International Conference on Intelligent Human-machine Systems & Cybernetics, IEEE, pp: 113 -116. (2014)
23. Zhou, S., Hui, Z.: Multi-document summarization using fully sparse topic models. Computer Engineering & Design, vol.35, no.3:1032-1036. (2014)
24. Yanga, G., Wenb, D., Kinshuk. B.: A novel contextual topic model for multi-document summarization. Expert Systems with Applications, vol.42, no.1: 1340–1352. (2015)
25. Wu, J., Shen, H., Li, Y.-D., Xiao, Z.-B., Lu, M.-Y., Wang, C.-L.: Learning a Hybrid Similarity Measure for Image Retrieval, Pattern Recognition, vol. 46, no. 11: 2927-2939. (2013)
26. Nenkova, L. Vanderwende.: The impact of frequency on summarization, Technical report, Microsoft Research. (2005)
27. Arora, R., Ravindran, B.: Latent Dirichlet Allocation Based Multi-document Summarization. Proceedings of the second workshop on Analytics for noisy unstructured text data,ACM, Singapore, pp: 91-97. (2008)
28. Chang,Y.-L., Chien, J.-T.: Latent Dirichlet Learning for Document Summarization. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Taipei, Taiwan, pp: 1689-1692. (2009).
29. Harabagiu, S., Lacatusu, F.: Generating single and multi-document summaries with gistexter. In Document Understanding Conferences. (2002)
30. Van Halteren, H.: Writing style recognition and sentence extraction. In U. Hahn and D. Harman (Eds.), Proceedings of the workshop on automatic summarization, pp:66–70. (2002)
31. Arora, R., Ravindran, B.: Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization. In ICDM 2008: Eighth IEEE International Conference on Data Mining, pp: 713–718. (2008)

**Liu Na** received M.S. and Ph.D. degree in Computer science and technology from Dalian Maritime University in 2004 and 2012 respectively. She is an Associate professor at Dalian Polytechnic University. Her research interests include data mining, machine learning. She is current with multi-document summarization and topic model.

**Di Tang** received her BS degree in Computer Science from Jilin University, China, in 1982. Currently she serves as a professor at the School of Computer and Information Technology, Liaoning Normal University, China. Her research mainly focuses on the computer graphics and computer vision.

**Lu Ying** received M.S. degree from Dalian University of Technology in 1999. She is a professor at Dalian Polytechnic University. Her research interests include Intelligent Information System, Network and multimedia technology applications.

**Tang Xiao-jun** received M.S. degree from Dalian University of Technology in 2002. She is a Associate professor at Dalian Polytechnic university, senior visiting scholar of West Oregon University. Her research interests include Intelligent Information System.

**Wang Hai-wen** received M.S. degree from Dalian University of Technology in 2004. He is an Associate professor at Dalian Polytechnic University. His research interests include Intelligent Information System.

*Received: August 15, 2014; Accepted: March 30, 2015.*