Aligning Gaussian-Topic with Embedding Network for Summarization Ranking

Linjing Wei[1,3], Heyan Huang[1,2], Yang Gao*[1,2], Xiaochi Wei[1,3], and Chong Feng[1,2]

1.Beijing Institute of TechnologyBeijing, China
2.Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
3.Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048,P.R.China

Abstract. Query-oriented summarization addresses the problem of information overload and help people get the main ideas within a short time. Summaries are composed by sentences. So, the basic idea of composing a salient summary is to construct quality sentences both for user specific queries and multiple documents. Sentence embedding has been shown effective in summarization tasks. However, these methods lack of the latent topic structure of contents. Hence, the summary lies only on vector space can hardly capture multi-topical content. In this paper, our proposed model incorporates the topical aspects and continuous vector representations, which jointly learns semantic rich representations encoded by vectors. Then, leveraged by topic filtering and embedding ranking model, the summarization can select desirable salient sentences. Experiments demonstrate outstanding performance of our proposed model from the perspectives of prominent topics and semantic coherence.

Keywords: Query-oriented summarization, Embedding model, Topic

1 Introduction

Sentence ranking, a vital part of extractive summarization, has been extensively investigated. In typical query-oriented summarization ranking, it often stresses two key points, one is to select the coherent salient sentences, and the other is to capture the desired topic content from the query. To this end, most traditional ranking models [7, 20, 26] utilize features (e.g., term frequency or position) to generate final summarization. Feature engineering largely determines the final summarization performance. Although they have acceptable performance and efficiency advantage, they lack of deep understanding semantics mechanism, therefore are unable to extract salient summaries.

Delighted by the successful embedding models, such as word2vec [18, 19] and PV [13], a great efforts [11, 12, 27, 4] have been conducted to construct summarization model. These methods utilize embeddings directly to calculate relevant

* Corresponding author.

sentences. Moreover, the word embeddings are generated based on language models that depend on local context. However, they lack of structures that embrace topic characteristics at global level. Vector-based representation only including local context cannot well represent sentence semantics, therefore, is not able to capture desired topical content for summaries.

Some works have realised the importance of combining topic analysis and word embedding approaches [17, 6, 3]. Despite of great success of combining topics in word embedding model for many text mining and NLP tasks, at sentence level, less work focuses on representing sentences with semantic topics in summarization tasks to our knowledge. To incorporate sentence embedding and topical analysis for enhancing summarization effectiveness and accuracy, two interesting questions are arising: (1) how to represent the consistent space for documents, sentences and words, that can facilitate to discover coherent semantics; (2) how to extract sentences from data collection that are dependent on the focused topics from the user query.

In order to tackle these problems, we propose a novel query-oriented multi-document summarization approach, called **G**aussian-**T**opic with **E**mbeddings Network for Summarization **F**iltering and **R**anking (GTEFR). Our proposed model is designed to leverage the topical aspects for continuous vector representations to facilitate sentence modelling for summarization. Gaussian distributions capture a notion of centrality in space, hence semantically related sentences are clustered together in the space. In order to solve the first problem, we encodes a prior preference for semantic topics and learn the topic distribution by incorporating a gaussian distribution into embedding network. In this way, the trained sentence embeddings are semantically rich and coherent. As to the second problem, we extract topic related sentences according to the learnt topic distributions of both query and candidate sentences.

In this paper, we consider the learning process as a nonlinear embeddings from content in terms of Gaussian mixture model (GMM) and a neural network framework. The words, sentences and documents are represented by the GMM of vectors. Particularly for sentence embeddings, the semantically related sentences are localized in the space. Therefore, the topic categories can represent sentences in an abstraction way, which can assist sentence embedding in a high level of guidance. Besides, every word and document have their own topic assignment to assure the learning consistency. During training process, all the vector representations are learnt upon the neural network and gradually updated according to the word sequential context and topic assignments of sentences and documents. Once this process ends, the system outputs coherent encoded vectors of words, sentences, documents and mixture topic distributions of them. With the result for topic distributions of queries and sentences, irrelevant sentences are filtered out by utilising the consistent topics of queries and sentences, then query-oriented summarization has advantages reflected by content coherence and relevancy. We conduct experiments to verify the effectiveness of the proposed model on a benchmark dataset, and quantitatively demonstrate

that our model outperforms those embedding models, the topic model, and the state-of-the-art topic-and-word embedding cooperation models.

The main contributions of our work include:

1. Aligning Gaussian-topic with embedding network is seamlessly integrated in the process of generating semantic sentences. Topics are jointly learnt in the GMM process as well as embedding learning process, which facilitate to aggregate topics accurately and capture intensive sentence semantics.
2. Representing queries and collection of sentences in terms of salient topics, in this way, desired topics are filtered especially based on the user specific needs. It enhances the topic coherence and relevance of summaries.
3. When referring to summarization tasks, we conduct experiments to demonstrate the effectiveness of our method for summarization ranking in real application.

The rest of this paper is organized as follows: Section 2 presents related work. We then proposed associated topics enhanced embedding model and summarization system in Section 3. Section 4 reports the experimental results and analysis. Finally, we conclude this paper in Section 5.

2 Related Work

Most existing extraction-based document summarization methods can be roughly divided into four categories, i.e., feature based method, deep learning based method, vector space based method and topic based method. Features such as term frequency [20, 26], cue words [14] and topic theme [10] are used to measure the importance of words. There are also some methods utilizing deep neural network, such as, LSTM [8], RBM [16]. Because our model is related to continuous distributed vector representation and topic model, we focus on these two categories.

Since Mikolov et al. [18] proposed the efficient word embedding method, vector space model has attracted a growth of attention. But to the best of our knowledge, only [11, 12] considered a direct summarization method using embeddings. Kågebäck et al. [11] proposed a summarization method, which maximized a submodular function defined by the summation of cosine similarity measure based on sentence embeddings. A summarization method based on document-level similarity [12] was proposed by Kobayashi et al, and they examined an objective function defined by a cosine similarity based on document embeddings instead of sentence embeddings.

Topic-based methods are widely applied in the summarization task. Parveen et al. [22] proposed an approach, which is based on a weighted graphical representation of documents obtained by topic modeling. Barzilay et al. [1] used the Hidden Markov Model to learn a latent topic for each sentence. They chose the "important" topics that had the high probability of generating summary sentences. The work proposed by Gupta et al. [9] measured topic concentration in a direct manner: a sentence was considered relevant to the query if it contained at

least one word from the query. While these work assume that documents related to the query only talk about one topic. Tang et al. [23] proposed a unified probabilistic approach to uncover query-oriented topics and four scoring methods to calculate the importance of each sentence in the document collection.

Although their success in the summarization task, semantic coherence and topic information are not encoded in these models. Sentence embedding with topic model has not previously been used in summarization tasks as far as we know, but there are same models incorporating vector representations and topics in other NLP tasks[17, 6, 25, 3]. Das et al. [6] developed a variant of LDA that operated on continuous space embeddings of words rather than word types to impose a prior, which helped topics to be semantically coherent. Liu et al [17] employed latent topic models to assign topics for each word in the text corpus, and learned topical word embeddings based on both words and their topics. Cao et al. [3] made use of the word embeddings available [18] and a neural network to explain the topic model on the embedding space. However, these methods learn the vector representations of words, rather than sentences, which is unfavourable for discovering comprehensive meaning on the summarization task. Our work is inspired by the recent neural probabilistic language models [25], which considered the ordering of words and the semantic meaning of sentences into topic modeling. While the work in [25] ignores the topic coherence among adjacent words. In our model, we add the sentence topic into context information which ensures topic coherence among adjacent words.

3 Proposed Model

In this section, we describe the details of the proposed method, which contains two stages, i.e. training embeddings and constructing summary based on the embeddings. Firstly, the embedding structure will be introduced, which seamlessly connects GMM and a embedding network. Then, we describe how to apply the proposed embedding method for summarization.

3.1 The embedding model GTE

The proposed model, aligning **G**aussian-**T**opic with **E**mbedding Network (GTE), is inspired by the recent work in learning vector representations of words and sentences using neural networks [17–19, 13, 25], we deploy a novel word prediction framework to exploit intensive sentence semantics. The model is built upon the GMM-based topic assumption of a continuous vector space. It is composed by two parts, the GMM centralised modelling and neural network learning framework, which is shown in Fig. 1. Given a document (sentence or word), the topical information is conveyed by its embedding vectors through the GMM. To be specific, the probability of words is calculated by the forward neural network and a nonlinear transformation function. In this section, we present the details of the proposed model GTE.

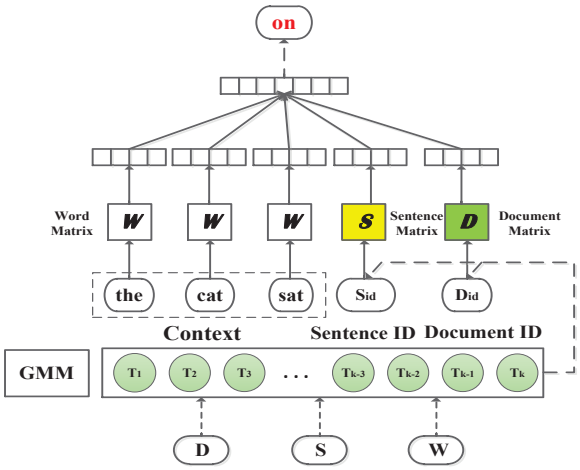


Fig. 1. The model architecture of the GTE model

Embedding process

Table 1 summarizes the notations. Given the collection of parameters of GMM,

Table 1. Notations

Symbol	Description	Symbol	Description
K	number of topics	$vec(d)$	vector of document d
W	words collection	$vec(s)$	vector of sentence s
S	sentence collection	$vec(w)$	vector of word w
D	document collection	ϕ	all vector representation set
M	number of documents	π	mixture weights of GMM
d_i	the ith document	μ	means of GMM
s_{ij}	the jth sentence in document d_i	Σ	covariance matrices of GMM
$\mathbf{T}_{d_i}$	the topic vector assigned to d_i	λ	The parameters collection of GMM
$\mathbf{T}_{s_{ij}}$	the topic vector assigned to s_{ij}		

we use

$$P(x|\lambda) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{1}$$

to represent the probability distribution for sampling a vector x from the Gaussian mixture model.

Given the model parameters λ and the vectors for documents, we can infer the posterior probability distribution of topics. For each document d_i, the posterior

distribution of its topic is

$$q(z(d_i) = z) = \frac{\pi_z N(vec(d_i)|\mu_z, \Sigma_z)}{\sum_{k=1}^{K} \pi_k N(vec(d_i)|\mu_k, \Sigma_k)} \tag{2}$$

Based on the distribution, the topic of document d_i can be vectorized as $[q(z(d_i) = 1), q(z(d_i) = 2), \cdots, q(z(d_i) = K)]$.

Similarly, for each sentence s_{ij} in the document d_i, the topic of sentence s_{ij} can be vectorized as $[q(z(s_{ij}) = 1), q(z(s_{ij}) = 2), \cdots, q(z(s_{ij}) = K)]$.

The basic idea of GTE is that we model one word as a prediction task based on word sequential context and topic assignments of sentences and documents. Given the Gaussian mixture model λ, the predicted process is described as follows:

1. For each document d_i in corpus D
   (a) Choose a topic $T_{d_i} \sim \pi := (\pi_1, \pi_2, \cdots, \pi_T)$
   (b) Choose a vector representation $vec(d_i) \sim N(\mu_{T_{d_i}}, \Sigma_{T_{d_i}})$
   (c) For each sentence s_{ij} in document d_i
      i. Choose a topic $T_{s_{ij}} \sim \pi := (\pi_1, \pi_2, \cdots, \pi_T)$
      ii. Choose a vector representation $vec(s_{ij}) \sim N(\mu_{T_{s_{ij}}}, \Sigma_{T_{s_{ij}}})$
      iii. For each word w_t in sentence s_{ij}
         A. Choose a topic $T_{w_t} \sim \pi := (\pi_1, \pi_2, \cdots, \pi_T)$
         B. Choose a vector representation $vec(w_t) \sim N(\mu_{T_{w_t}}, \Sigma_{T_{w_t}})$
2. For t-th word $vec(w_t)$ in sentence s_{ij}
   (a) Predict w_t according to the documents vector $vec(d_i)$ and topic $\mathbf{T}_{d_i}$, the current sentences vector $vec(s_{ij})$ and topic $\mathbf{T}_{s_{ij}}$, as well as at most m previous words in the same sentence.

Given the t-th location in j-th sentence, we represent its word realization by w_t, the objective of GTE is to maximize the probability

$$G_t = P(w_t|d_i, s_{ij}, w_{t-m}, \cdots, w_{t-1}) \tag{3}$$

As aforementioned, the assumption is that the word is predicted by the representation of the word's sentence and document as well as their assigned topics. Besides, as we all know, dot-product represents similarity of two vectors, so maximizing the similarity is utilized to guide vector embedding and topic mutually. Thus, we obtained the following objective function

$$G_t = \sigma(\mathbf{T}_{d_i} vec(d_i)^T + \mathbf{T}_{s_{ij}} vec(s_{ij})^T + \sum_{n=1}^{m} \mu_n^{w_t} vec(w_{t-n})^T) \tag{4}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ and $\mu_n^{w_t} \in R^V$ is parameter of the model.

Combining the equations above, we use GMM as a prior and generate the next word by a sigmoid conditional distribution. So the log-likelihood of the generative model can be described as

$$J_t = log(P(\phi|\lambda)) + log(\sigma(\mathbf{T}_{d_i} vec(d_i)^T + \mathbf{T}_{s_{ij}} vec(s_{ij})^T + \sum_{n=1}^{m} \mu_n^{w_t} vec(w_{t-n})^T)) \tag{5}$$

Estimating model parameters
The model parameters $\{\lambda, \mu_n^{w_t}, \phi\}$ are estimated by maximizing the likelihood of the generative model. A two-phase iteration process is conducted, as shown in Algorithm 1. Given $\{\mu_n^{w_t}, \phi\}$, the parameters of the Gaussian mixture model $\lambda = \{\pi_k, \mu_k, \Sigma_k\}$ are estimated by Expectation Maximization (EM) algorithm. Given λ, stochastic gradient descent (SGD) is adopted to find the optimized result.

Fig. 2 shows a small example of the results after estimating model parameters and learning. We list several vector and topic representations of documents, sentences as well as words on document collection D301 of DUC2005. As shown in Fig. 2, the vector and topic representations of document d_1 and d_2, coming from document collection D301, can be learned since the GTE model is performed on multiple documents of the D301. Likewise, the proposed model also learn vector and topic representations of s_{11} and s_{21}, w_1 and w_2. What's more, w_1 and w_2 are selected from s_{11} and s_{21}, respectively.

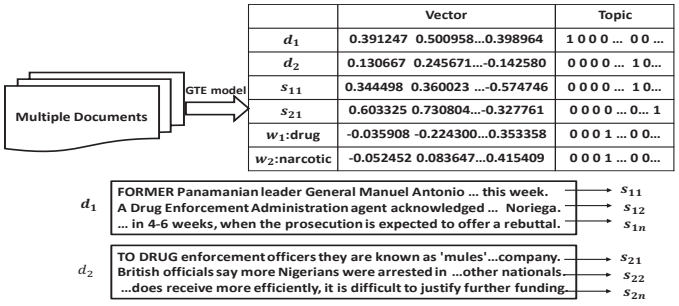


Fig. 2. The examples of vectors and topics

3.2 The summarization model GTEFR

We regard the informative results from the GTE model on different granularities of topical distributions, sentence and word embeddings as an exhaustive mining of the whole collection of documents. As a result, it will strongly support our summarization ranking system.

Fig. 3 shows the graphical structure of the summarization model. In order to extract desired topic content and salient sentences, we construct two submodules i.e. sentence filtering and sentence ranking. Sentence filtering facilitates to filter out those irrelevant sentences to the user desired content from topic perspective. while for ranking, the system considers the performance of the summary at two different levels simultaneously i.e. word and sentence level.

Sentence Filtering
Sentence filtering in our summarization framework aims to learn the salient topics for updating a query-focused sentence collections and filter out those

Algorithm 1

Input:

Documents D, sentences S, $|W|$ words contained in dictionary W, the learning rate α, the dimension of the vector V and the number of topics K.

Output:

Topic representations $\mathbf{T}_w$, $\mathbf{T}_s$ and $\mathbf{T}_d$. vector representations $vec(w)$, $vec(s)$ and $vec(d)$.

1: Randomly initialize parameters.

2: Fixing parameters $\mu_n^{w_t}$ and ϕ, run the EM algorithm:

E-Step:

$\gamma(i,k) = \frac{\pi_k N(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j,\Sigma_j}$

M-Step:

$N_k = \sum_{i=1}^{N} \gamma(i,k)$

$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i$

$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T$

$\pi_k = \frac{N_k}{N}$

3: Fixing parameters λ, run the SGD algorithm:

For each document d_i

For the t-th word w_t of sentence s_{ij}

Update the vector of document d_i, s_{ij} and w_t

Update the factor of influence $\mu_n^{w^t}$

$vec(d_i) \leftarrow vec(d_i) + \alpha \frac{\partial J_t(\mu_n^{w_t}, \phi)}{\partial vec(d_i)}$

$vec(s_{ij}) \leftarrow vec(s_{ij}) + \alpha \frac{\partial J_t(\mu_n^{w_t}, \phi)}{\partial vec(s_{ij})}$

$vec(w_t) \leftarrow vec(w_t) + \alpha \frac{\partial J_t(\mu_n^{w_t}, \phi)}{\partial vec(w_t)}$

$vec(\mu_n^{w_t}) \leftarrow vec(\mu_n^{w_t}) + \alpha \frac{\partial J_t(\mu_n^{w_t}, \phi)}{\partial vec(\mu_n^{w_t})}$

End For

End For

irrelevant sentences quickly and efficiently. Specifically, the GTE process outputs topic distributions of query and each sentence. With obtaining the topic distribution of one query, we can rank the topic of the query in descending order. Subsequently, the salient topics set $salient(T)$ of query can be defined as $salient(T) = \{T_k | k > p, k \in \{1, 2, \cdots, K\}\}$, where T_k denotes the kth topic after ranking the topic distribution and p is a parameter in our model.

Similarly, with obtaining the topic distribution of one certain sentence, we select the top p sentence topics. If one of these topic is contained in the salient topics set $salient(T)$, the sentence will be appended to the new query-focused sentence collections.

Sentence Ranking

As obtaining word and sentence embedding after the GTE process, we use a weighted sum of all scores of the two different levels (word and sentence level) to measure the importance of sentences in the new query-focused sentence

collections. Then, the sentence score is described as follows:

$$Score(s) = \underbrace{\beta \sum_{t=1}^{n_w} TF(w_t) + \gamma \sum_{t=1}^{n_w} Sim(vec(w_t), vec(Q_1))}_{word \quad level} + \underbrace{\delta Sim(vec(s), vec(Q_2))}_{sentence\ level} \tag{6}$$

where n_w represents the number of words in sentence s. Q_1 is the word-based query[1], which is several key words and contributions to focus the user interests roughly. Q_2 represents the sentence-based query[2], which is a complete sentence and relates to a in-depth and comprehensive aspect of the subject. $Sim()$ represents the function of similarity, and we use cosine similarity in this paper. β, γ and δ are parameters in our summarization model.

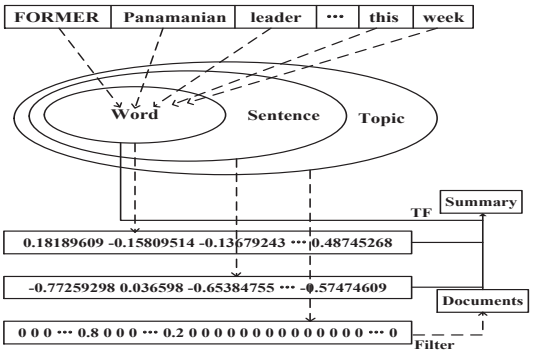


Fig. 3. The graphical structure of the summarization model

At word level, the importance of words is calculated by a weighted sum of term frequency and the relevant score between sentence s_{ij} and word-based query. Vector representation of a sentence retains the coherently semantic information. The most coherent and important sentences are selected by the cosine similarity which represents the similarity between each sentence and query.

4 Experiments

In this section, we present the evaluation of the proposed approach. The hypothesis is that our model is effective in two aspects as follows: 1) Incorporating vector representations of sentences and topics is contributed to capture the coherent semantic meaning and focused topical content, 2) our model is effective in summarization tasks.

[1] In DUC, the word-based query is also called "title", such as "New hydroelectric projects"

[2] In DUC, the sentence-based query is also called "narrative", such as "What hydroelectric projects are planned or in progress and what problems are associated with them?"

4.1 Data Sets

In this study, we use the standard summarization benchmark DUC2005[3] for evaluation. DUC2005 contains 50 query-oriented summarization tasks. For each query, a relevant document cluster is assumed to be "retrieved", which contains 25-50 documents. Thus, the task is to generate a summary from the document cluster for answering the query[4]. The length of a result summary is limited by 250 tokens (whitespace delimited).

For the benchmark data sets, we preprocessed each document by(a) removing stopwords; (b) removing words that appear less than one time in the corpus; and (c)downcasing the obtained words.

4.2 Evaluation Measures

We conducted evaluations by ROUGE [15] metrics. The measure evaluates the quality of the summarization by counting the number of overlapping units, such as n-grams. Basically, ROUGE-N is an n-gram recall measure. Among the evaluation methods implemented in Rouge, Rouge-1 emphasises on the occurrence of the same words between candidate summary and reference summary, while Rouge-2 and Rouge-SU4 focus more on the readability of the candidate summary. We use these three metrics in the experiment.

4.3 Baseline Models

We compare the GTE model with several query-focused summarization methods.

1. **TF:** this model uses term frequency [20] for scoring words and sentences.
2. **Lead:** take the first sentences one by one from the document in the collection, where documents are ordered randomly. It is often used as an official baseline of DUC.
3. **Avg-DUC05:** average system-summarizer performance on DUC2005.
4. **LDA:** this method uses Latent Dirichlet Allocation[2] to learn the topic model. After learned the topic model, we give max score to the word of the same topic with query. The reader can refer to the paper [23] for the details.
5. **LLRSum:** This system [5] employs a log-likelihood ratio (LLR) test to select topic words. The sentence importance score is equal to the number of topic words divided by the number of words in the sentence.
6. **SNMF:** this system [24] is for topic-biased summarization. it used symmetric non-negative matrix factorization (SNMF) to cluster sentences into groups, then selected sentences from each group for summary generation.
7. **Word2Vec:** the vector representations of words can be learned by Word2Vec [18, 19] models. We use the vectors and calculate the three features, where the sentence-level representations is calculated by using a average of all word embeddings in the sentence.

[3] http://duc.nist.gov/data.html

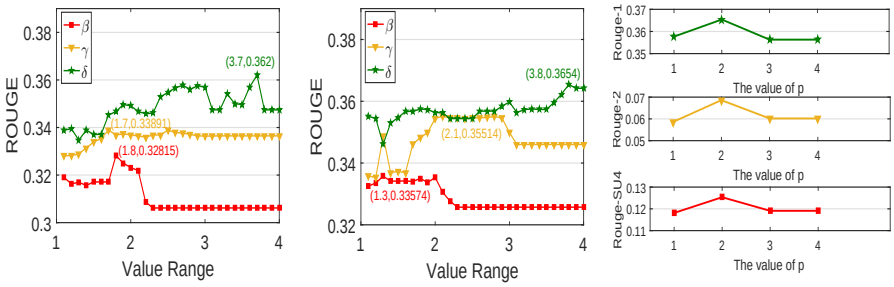
[4] In DUC, the query is also called "narrative" or "topic"

8. **PV:** PV [13] learns sentence vectors based on Word2Vec Model. Thus, we use the same parameters as that in our approach to calculate the scores of sentences without word similarity.
9. **TWE:** TWE [17] learns topical word embeddings based on both words and their topics. The three features are calculate with the same parameters, where the sentence-level representations is calculated by using a average of all word embeddings in the sentence.
10. **GTER:** Comparing with GTEFR, we implement the summarization utilize sentence ranking submodule without sentence filtering.

4.4 Implementation details

In the training stage, the learning rate α is set to 0.026 and gradually reduced to 0.0001. For each word, at most m = 6 previous words in the same sentence is used as the context. The word vector size is set to the same as the number of topics V = K = 100. These parameters (m, V, K) are empirically set in our experiment.

In the summarization stage, we use the MERT [21] to tune parameters for GTER and GTEFR, which is showed in Fig.4(a) and Fig.4(b). γ, δ and ρ are tuned from 1 to 10, with the step size of 0.1. What's more, the parameter p is tuned from 1 to 4, which is showed in Fig.4(c), and finally we set $p = 2$. As mentioned above, a relevant document cluster is assigned to a query. Thus, documents related to the query talk about a main topic. Then our experiment show that a document cluster averagely talks about two topics. Similar result has been mentioned in Tang et al. [23], who think a document cluster talks about multiple topics.



(a) The process of the MERT for GTER. (b) The process of the MERT for GTEFR. (c) The ROUGE of the topic number p.

Fig. 4. (a) the MERT for GTEFR and (b) the MERT for GTEFR. (c) ROUGE via tuning the parameter p

All experiments were carried out on a CentOS 7.2 Server with four Dual-Core Intel Xeon processors (2.6 GHz) and 8GB memory. It took about 3 days for estimating the GTE model. There is a problem of high time complexity in training process. But we all know the training process is offline, and then the efficiency is not primary issue.

4.5 Experimental Results

In this subsection, we give the results of the experiments and the analysis.

Table 2. Results of the baseline methods and the GTE model on DUC2005

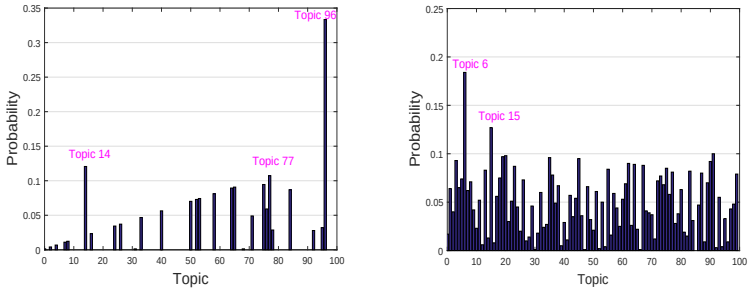
Method	Rouge-1	Rouge-2	Rouge-SU4
TF	0.3356	0.0580	0.1107
Lead	0.3354	0.0565	0.1102
Avg-DUC05	0.3434	0.0602	0.1148
LDA	0.3170	0.0533	0.1150
LLRSum	0.3363	0.0599	0.1166
SNMF	0.3501	0.0604	0.1229
Word2Vec	0.3459	0.0548	0.1154
PV	0.3541	0.0614	0.1186
TWE	0.3505	0.0606	0.1202
GTER	0.3600	0.0633	0.1231
GTEFR	**0.3620**	**0.0636**	**0.1244**

The results for the proposed model and baseline models are reported in Table 2. As shown in the table, the scores in bold are the highest ones in the column. It can be observed that our model gives the best summary compare to any other method in ROUGE measurements, which strongly testifies the effectiveness of the proposed summarization model. Meanwhile, our system significantly outperforms these methods based on features (i.e. TF and Lead). It is because that our method wisely considers deep semantics and desired topic content.

Impact by topics From Table 2, we can also obtain that, comparing with the embedding model like Word2Vec and PV, the GTER and GTEFR are always outperforming in all three ROUGE measurements. As aforementioned, Word2Vec and PV learn vector representations mainly based on the local context information, none of them take into consideration of global topic information that contributes to capture coherent semantics and obtain the desired topic content. This is main reason why our models are outperforming those state-of-the-arts high-quality embedding methods for summarization. Besides, the performance of the GTEFR model exceeds the GTER model, which also strongly demonstrates the effectiveness of sentence filtering based on topics for obtaining desired topics.

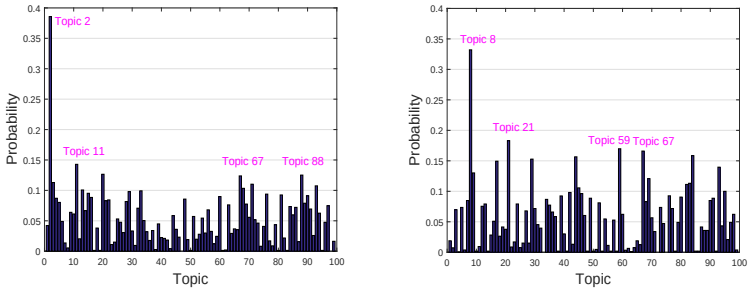
To reveal the reason why the topic-based approaches (e.g., GTEFR) can outperform the word-based approaches and the embedding-based approaches (e.g., TF, Word2Vec and PV), a topical analysis was performed on the document collections. As the Fig.5 shows, we calculated the probability distribution of each topic, since for a certain sentence, its topic can be learned by the GTE model. From the Fig.5(a), we see that there is a major topic (Topic 96, talking about new hydroelectric projects), but still same information is captured by the other topics (e.g., ♯14 and ♯77). While word-based approaches and embedding-based approaches can not learn the latent topic structure. It indicates that the proposed model can accelerate to focus the salient topics and further enhance summarization accuracy. Besides, topic coverage can be guaranteed, which facilitates to meet the summary need. We have also found that topics are unbalanced

Fig. 5. Topic distribution for collections. The x axis denotes topics and the y axis denotes the occurrence probability of each topic in collections



(a) Topic distribution for collection D307.

(b) Topic distribution for collection D311.



(c) Topic distribution for collection D313.

(d) Topic distribution for collection D389.

distributed in a collection. However, viewing the Fig.5 integratedly, topics are covered by all collections, which denotes the data is balanced distributed rather than tend to a certain category. what's more, each document can be distinguished by topic distribution. Hence, aligning Gaussian-topic effectively boosts to extract topic information for summarizer.

Impact by sentence modelling Our proposed model performs better than the TWE model, and the PV model performs better than the Word2Vec model. These results suggest that sentence-level methods can facilitate to discover more semantic contents than the word-level method. On the other hand, our model and the PV model perform better than the Word2Vec, especially in Rouge-2 and Rouge-SU4. As mentioned above, Rouge-2 and Rouge-SU4 focus more on the readability of the candidate summary. Because our model considers vector representation of sentence which retains the complete and coherent semantic content. What's more, our model is better than all the topic-based baseline models, such as LDA, LLRSum and SNMF. LDA aims to cover the main topics of documents and LLRSum pays more attention to find topic word. Although SNMF select sentences from each topic group, it ignores the importance of salient topics. More importantly, none of them take into consideration of sentence coherence. All of the aforementioned analysis of results prove that encoding sentences and topics into the same process can facilitate to discover coherent and intensive semantics, especially for summarization tasks.

Table 3. Contribution analysis of each feature to the final summary generation

Method			Rouge-1	Rouge-2	Rouge-SU4
TF-IDF	word_sim	sen_sim			
√	√		0.3228	0.0500	0.1107
	√	√	0.3308	0.0528	0.1146
√		√	0.3562	0.0641	0.1198
√	√	√	0.3654	0.0686	0.1254

In order to explore impact of the designed measure of word and sentence similarity deeply, we remove the part of features and keep the rest of feature parameters consistent. Table 3 presents the contribution analysis of each feature to the final summary generation. From the table, we can observe that sentence embedding achieves the best performance for summarization. What's is more, the performance of word embedding is comparable to general baseline models. we can imply that sentence similarity computation by our proposed sentence embedding plays the dominant impact for the summary and aligning Gaussian-topic with the neural network is effectively boosting to capture intensive sentence semantics.

5 Conclusion

A new model is proposed to leverage topics and vectors to capture complete sentence meaning. In the model, the topical embedding, document, sentence and word embeddings are jointly learnt through the seamlessly combinative GMM model and embedding network, and then supportively adapted to extract the more semantic relevant, coherent and topical focused summaries. Comparing to all baselines summarization approaches, our proposed model outperforms in all three ROUGE measurements. Besides, the proposed approach is quite flexible, which can also be applied to other areas, such as question answering and text classification.

Acknowledgments

The work was supported by National Nature Science Foundation of China (Grant No.61602036), National Basic Research Program of China (973 Program, Grant No.2013CB329303), Beijing Advanced Innovation Center for Imaging Technology (BAICIT-2016007).

References

1. Barzilay, R., Lee, L.: Catching the drift: Probabilistic content models, with applications to generation and summarization. Computer Science (2004)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR (2003)
3. Cao, Z., Li, S., Liu, Y., Li, W., Ji, H.: A novel neural topic model and its supervised extension. In: Proceedings of AAAI'15 (2015)

4. Chen, K.Y., Liu, S.H., Wang, H.M., Chen, B., Chen, H.H.: Leveraging word embeddings for spoken document summarization. Computer Science (2015)
5. Conroy, J.M., Schlesinger, J.D., O'Leary, D.P.: Topic-focused multi-document summarization using an approximate oracle score. In: Proceedings of ACL'06 (2006)
6. Das, R., Zaheer, M., Dyer, C.: Gaussian lda for topic models with word embeddings. In: Proceedings of ACL'15 (2015)
7. Galley, M.: A skip-chain conditional random field for ranking meeting utterances by importance. In: Proceedings of EMNLP'07 (2006)
8. Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., Heck, L.: Contextual lstm (clstm) models for large scale nlp tasks
9. Gupta, S., Nenkova, A., Jurafsky, D.: Measuring importance and query relevance in topic-focused multi-document summarization (2007)
10. Harabagiu, S., Lacatusu, F.: Topic themes for multi-document summarization. In: Proceedings of SIGIR'05 (2005)
11. Kågebäck, M., Mogren, O., Tahmasebi, N., Dubhashi, D.: Extractive summarization using continuous vector space models. In: Proceedings of EACL'14 (2014)
12. Kobayashi, H., Noguchi, M., Yatsuka, T.: Summarization based on embedding distributions. In: Proceedings of EMNLP'15 (2015)
13. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. Computer Science (2014)
14. Lin, C.Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In: Proceedings of COLING'00 (2000)
15. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of ACL'03 (2003)
16. Liu, Y.: Query-oriented multi-document summarization via unsupervised deep learning. In: Proceedings of AAAI'12 (2012)
17. Liu, Y., Liu, Z., Chua, T.S., Sun, M.: Topical word embeddings. In: Proceedings of AAAI'15 (2015)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computer Science (2013)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (2013)
20. Nenkova, A., Vanderwende, L., Mckeown, K.: A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In: Proceedings of SIGIR'06 (2006)
21. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of ACL'03 (2003)
22. Parveen, D., Ramsl, H., Strube, M.: Topical coherence for graph-based extractive summarization. In: Proceedings of EMNLP'15 (2015)
23. Tang, J., Yao, L., Chen, D.: Multi-topic based query-oriented summarization. In: Proceedings of SDM'09 (2009)
24. Wang, D., Li, T., Zhu, S., Ding, C.: Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In: Proceedings of SIGIR'08 (2008)
25. Yang, M., Cui, T., Tu, W.: Ordering-sensitive and semantic-aware topic modeling. In: Proceedings of AAAI'15 (2015)
26. Yih, W.T., Goodman, J., Vanderwende, L., Suzuki, H.: Multi-document summarization by maximizing informative content-words. In: Proceedings of IJCAI'07 (2007)
27. Yin, W., Pei, Y.: Optimizing sentence modeling and selection for document summarization. In: Proceedings of IJCAI'15 (2015)