Topical Sentence Embedding for Query Focused Document Summarization

Yang Gao
Beijing Institute of Technology (BIT);
Beijing Engineering Research
Center of High Volume Language
Information Processing and
Cloud Computing Applications
gyang@bit.edu.cn

Linjing Wei
BIT; Beijing Advanced Innovation Center for
Imaging Technology, Capital Normal University
weilinjing@bit.edu.cn

Heyan Huang
BIT; Beijing Engineering Research
Center of High Volume Language
Information Processing and
Cloud Computing Applications
hhy63@bit.edu.cn

Qian Liu
BIT; Beijing Advanced Innovation Center for
Imaging Technology, Capital Normal University
liuqian2013@bit.edu.cn

Abstract

Distributed vector representation for sentences have been utilized in summarization area, since it simplifies semantic cosine calculation between sentence to sentence as well as sentence to document. Many extension works have been done to incorporate latent topics and word embedding, however, few of them assign sentences with explicit topics. Besides, much sentence embedding framework follows the same spirit of prediction task about a word in the sentence, which omits the sentence-to-sentence coherence. To address these problems, we proposed a novel sentence embedding framework to collaborate the current sentence representation, word-based content and topic assignment of the sentence to predict the next sentence representation. The experiments on summarization tasks show our model outperforms state-of-the-art methods.

In: Proceedings of IJCAI Workshop on Semantic Machine Learning
(SML 2017), Aug 19-25 2017, Melbourne, Australia.

1 Introduction

Text summarization is an important task in natural language processing, which is expected to understand the meaning of the documents and then produce a coherent, informative but brief summarization of the original document with in a limited length. The main approaches of text summarization can be divided into two categories: extractive and generative. Most extractive summarization systems extract parts of the document (a few sentences or a few words) that are deemed interesting by some metric (i.e., inverse-document frequency) and join them to form a summary. Conventionally, selecting sentences rely on feature engineering approach in terms of extracting surface feature statistics (i.e., TFIDF cosine similarity) to compare with query and document representation.

Recently, distributed vector semantic representation for words and sentences have achieved overwhelming success in summarization area [KMTD14, KNY15, YP15], since it converts high-dimensional and sparse linguistic data into a controllable and dense dimension of semantic vectors. It becomes more straightforward for generic summarization to compute similarity (or relevance to some extents) and facilitates semantic calculation. Delighted by the successful word2vec model [MCCD13, MSC$^+$13], Paragraph Vector (PV) [LM14] model (i.e., the paragraph can be sentence, paragraph or document) also contributes to predict the next word given sequential word context and the current paragraph representation. It inherits the semantic representation and its efficiency, further captures the word order for sentence representation. Moreover, the sentence vector can benefit summaries since it directly characterises the relevance between queries and candidate sentences.

However, most of the sentence embedding models [LM14, YP15] are trained as the prediction task about a word in the sentence. In these models, sentences are independently learnt via their local word content but often omit the coherent relationship between sentences. Summarization system focuses more on comprehensive attributes of sentences, such as sentence coherence, sentence topic, sentence representation and so on. Utilizing the conventional sentence vectors may neglect the coherence between candidate sentences as well as sentence topics. Although, models incorporating topic and word embedding models, such as TWE [LLCS15], have achieved successful results in some NLP tasks, at sentence level, very few work focuses on representing sentences with topics. For example, given a user's query that emphasises on possible plans, progress and problems with *hydroelectric projects*. The query contain complex topics like "plans", "progress", "problems" and "hydroelectric projects". Nevertheless, normal vector-based models can retrieve those relevant sentences that only emphasis on one or two aspects of the query. It is problematic to capture all the aspects of the query .

In order to tackle the problems, we propose a novel sentence embedding learning framework to enhance sentence representation by incorporating multi-topic semantics for summarization task, called Topical Sentence Embedding (TSE) model. Gaussian distributions are utilised to model mixtured centralities of the embedding space, which capture a prior preference of topic for sentence prediction. In addition, instead of training to predict words in the document, our proposed model represents one sentence by predicting the next sentence via jointly training the words in the current sentence and the topic of the sentence.

The rest of this paper is organized as follows. Section 2 summarizes the basic methods of embedding models and summarization systems. We then introduce a newly summarization framework in Section 3, especially in Section 3.2, the novel TSE model is proposed. Section 4 reports the experimental results and corresponding analysis. Finally, we conclude the paper.

2 Background and Related Work

We firstly introduce the Word2Vec and the PV model to investigate the basic framework of training embedding model for words and sentences.

Word2Vec:

The basic assumption behind Word2Vec [MCCD13] is that the representation of co-occurred words have the similar representation in the semantic space. To this target, a sliding window is employed on the input text stream, where the central word is the target word and others are contexts. Word2Vec method contains two models: CBOW and Skip-gram model. CBOW aims at predicting the target word using the context words in the sliding window. The objective of CBOW is to maximize the average log probability,

$$L = \frac{1}{D}\sum_{i=1}^{D} \log Pr(w_i \mid C; W). \quad (1)$$

where, w_i is the target word, C is the word contexts and W is is word matrix, D is the corpus size. Different from CBOW, Skip-gram aims to predict context words given the target word. We ignore the details of this approach here.

Paragraph Vector (PV):

It [LM14] is an unsupervised algorithm that learns fixed-length semantic representations of variable-length of texts, which follows the same predicting task with Word2Vec. The only change is the concatenate vector constructed from W and S, where S is sentence matrix instead of individual W. The PV model is a strong alternative sentence model, and it is widely applied in learning representations for sequential data.

Work on extractive summarization spans a large range of approaches. Most existing systems [Gal06, YGVS07] use rank model to select the sentences with highest scores to form the summarization. However, multi-document texts often describe one central topic and some sub-topics, which cannot be described only depending on ranking model. Then we focus on how to rank the sentences and collaborate topic coverage.

A variety of features were defined to measure the relevance, including TF-IDF cosine similarity [NVM06, YGVS07], cue words [LH00], topic theme [HL05], and WordNet similarity [OLLL11], etc. However, these features usually suffer from lacking of deep understanding semantics mechanism, which fail to meet the query need. Since Mikolov et al. [MCCD13] proposed the efficient word embedding method, there is a surge of works [LM14, LLCS15] focusing on embedding models for capturing the linguistic regularities. Embedding models [KMTD14, KNY15, YP15, CLW$^+$15] for words and sentences also have encouraged summarization tasks from the perspective of semantic relevance computing, such as DocEmb and CNNLM. However, aforementioned methods usually reward semantic similarity without considering of topic coverage, which fail to meet the summary need.

Topic-based methods have been proved their successes for summarization. Parveen et al. [PRS15] proposed an approach, which is based on a weighted graphical representation of documents obtained by topic modeling. [GNJ07] 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. [TYC09] 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. Wang et al. [WLZD08] propose a new multi-document summarization framework (SNMF) based on sentence-level semantic analysis and symmetric non-negative matrix factorization. The symmetric matrix factorization has been shown to be equivalent to normalized spectral clustering and is used to group sentences into clusters. Futhermore, several approaches incorporate vector representations with topics , such as NTM [CLL$^+$15], TWE [LLCS15] and GMNTM [YCT15], have collaborated both benefits of semantic representation and classified topics. This motivates us to investigate the cooperation models for summarization system.

3 The Framework for Query-focused Summarization

Extracting salient sentences is the main task in this study. At sentence level, the sentence embedding and sentence ranking are utilised to enable sentence relevance to the user queries and extract salient summaries.

3.1 The Proposed TSE Model

Inheriting the superiority of the PV model that constructs a continuous semantic space, the novel architecture of learning sentence representation, called TSE model, as shown in the Figure 1.

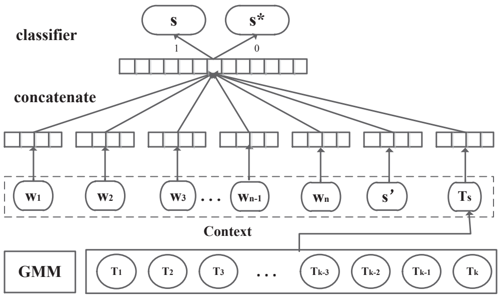

Figure 1: The structure of the proposed TSE model

Topic Vectorization by GMM

Let K represent the number of topics, V is the size of vector, and W represent word dictionary. S denotes the sentence collection, in which s is one of the sentences. Let $vec(T_s)$ be the topic vector of sentence s. The vectors of sentences and words are represented as $vec(s) \in R^V$ and $vec(w) \in R^V$. $\pi_k \in R$, $\mu_k \in R^V$, $\Sigma_k \in R^{V \times V}$ and $\sum_{k=1}^{K} \pi_k = 1$ are denoted as mixture weights, means and covariance matrices, respectively. The parameters of the GMM are collectively represented by $\lambda = \{\pi_k, \mu_k, \Sigma_k\}$, where $k = 1, \cdots, K$. Given the collection of parameters, we use

$$P(x|\lambda) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{2}$$

to represent the probability distribution for sampling a vector x from the GMM.

Subsequently, we can infer the posterior probability distribution of topics. For each sentence s, the posterior distribution of its topic is

$$q(z_s = k) = \frac{\pi_z N(vec(s)|\mu_z, \Sigma_z)}{\sum_{k=1}^{K} N(vec(s)|\mu_k, \Sigma_k)} \tag{3}$$

Based on the distribution, the topic of sentence s can be vectorized as $vec(T_s) = [q(z_s = 1), q(z_s = 2), \cdots, q(z_s = K)]$.

Generative Sentence Embedding

The assumption of the TSE is that sentences are coherent and associated with their neighbours. Consequently, we model one sentence as a prediction task based on semantic structure of the previous sentences. The semantic is represented by collaborating sentence topic, sentence representation and its content. The Negative Sampling (NEG) method is applied in [MCCD13] which is an efficient approximation method. Therefore, we carry on the similar estimation schema in our model.

Definition 1. Label $l^{\widetilde{s}}$: *A label of sentence* $\widetilde{s}$ *is 1 or 0. The label of positive sample is 1, the label of negative samples are 0.* For $\forall \widetilde{s} \in S$,

$$l^s(\widetilde{s}) = \begin{cases} 1, \widetilde{s} = s; \\ 0, \widetilde{s} \neq s; \end{cases} \tag{4}$$

Let X_s be a concatenation of given information of current sentence for predicting the next sentence, s, s' be the current sentence. $X_s = vec(T_{s'}) \oplus vec(s') \oplus vec(w_1) \oplus, \cdots, \oplus vec(w_m)$. We incorporate the vectors as the input, which includes topics, sentence embedding, and its content of words.

Given the collection S, we show how to learn representation of sentences and topics. In this paper, we concentrate to exploit the latent relationship between sentences. Subsequently, the target sentence s is predicted purely by the information from previous sentence, namely X_s. So the objective of TSE is to maximize the probability

$$G = \prod_{s \in S} g(s) = \prod_{s \in S} \prod_{u \in \{s \cup s^-\}} p(u|X_s) \tag{5}$$

Instead of using softmax function as prediction probability, we directly use its negative sampling approximation. The prediction objective function of sentence s is g(s)=$\prod_{s \in S} p(u|Xs)$, and the probability function is represented as follows

$$p(u|X_s) = \begin{cases} \sigma(X_s^T \theta^u), l^s(\widetilde{u}) = 1 \\ 1 - \sigma(X_s^T \theta^u), l^s(\widetilde{u}) = 0 \end{cases} \tag{6}$$

or write as a whole

$$p(u|X_s) = [\sigma(X_s^T \theta^u)]^{l^s(\widetilde{u})} \cdot [1 - \sigma(X_s^T \theta^u)]^{1 - l^s(\widetilde{u})} \tag{7}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ and $\theta^u \in R^V$ is the parameter of X_s.

The objective function is taken log-likelihood and defined as

$$\mathcal{L} = \sum_{s \in S} l^s(u) \log[\sigma(X_s^T \theta^u)] + (1 - l^s(u))(n\mathbb{E}(s^* \sim \mathbb{N}(S))) \log[1 - \sigma(X_s^T \theta^u)] \quad (8)$$

where $n\mathbb{E}(\cdot)$ is number of n negative samples as Definition 1, and n is set to 10 empirically. Considering convenience in estimation, we rewrite the final objective function as

$$\mathcal{L}(s, u) = l^s(u) \cdot \log[\sigma(X_s^T \theta^u)] + [1 - l^s(u)] \cdot \log[1 - \sigma(X_s^T \theta^u)] \quad (9)$$

Parameters Estimation

The parameters $\{\lambda, \theta^u, X_s\}$, where $\lambda = \{\pi_k, \mu_k, \Sigma_k\}$ are estimated by maximizing the likelihood of the objective function jointly. A two-phase iteration process is conducted.

Given $\{\theta^u, X_s\}$, stochastic gradient descent (SGD) is adopted in updating parameters of the GMM. Given λ, the gradient of θ^u is calculated using the back propagation based on the objective in Eq. 9.

3.2 Sentence Ranking

Sentence ranking aims to measure the relevant sentences with consideration of query information. In this paper, relevance ranking of sentences primarily relys on semantic vector-based cosine similarity [KMTD14] that is a promising measure to compute relatedness for summarization. Additionally, statistics features (i.e., TFIDF score [NVM06]). In summary, the ranking score is formulated as:

$$Score(S) = \alpha \sum_{t=1}^{n_w} TFIDF(w_t) + \beta sim(vec(s), vec(Q)) + \gamma sim(vec(T_s), vec(T_Q)) \quad (10)$$

where Q is the query, $sim(\cdot)$ represents the function to compute similarity, and we use cosine similarity in this paper. α, β and γ are parameters in the summarization system.

4 Experiments

In this section, we present experiments to evaluate the performance of our method in query focused multi-document summarization task.

4.1 Dataset and Evaluation Metrics

In this study, we use the standard summarization benchmark DUC2005 and DUC2006[1] 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. DUC2006 contains 50 query-oriented summarization tasks as well and each query contains 25 documents. Thus, the task is to generate a summary from the document cluster for answering the query[2]. The length of a result summary is limited by 250 words.

We conducted evaluations by ROUGE [LH03] metrics. The measure evaluates the quality of the summarization by counting the number of overlapping units, such as n-grams. Basically, ROUGE-N is n-gram recall measure.

4.2 Baseline Models and Settings

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

- **TF-IDF:** this model uses TF-IDF [NVM06] for scoring words and sentences.
- **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.
- **LDA:** this method uses Latent Dirichlet Allocation[BNJ03] 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 [TYC09] for the details.
- **SNMF:** this system [WLZD08] is for topic-biased summarization. It utilised non-negative matrix factorization (SNMF) to cluster sentences and from which selected multi-coverage summary sentences.
- **Word2Vec:** the vector representations of words can be learned by Word2Vec [MCCD13, MSC$^+$13] models. The sentence representation is calculated by using an average of all word embeddings in the sentence.
- **PV:** PV [LM14] learns sentence vectors based on Word2Vec Model. Thus, we use the same parameters as that in our approach to calculate the scores of sentences.
- **TWE:** TWE [LLCS15] employs LDA to refine Skip-gram model. It learns topical word embeddings based on both words and their topics. The sentence representation is calculated by using an average of all word vectors in the sentence.

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

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

Table 1: Overall ROUGE evaluation (%) of different models for DUC2005 and DUC 2006

Method	DUC2005		DUC2006	
	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
LEAD	29.71	4.69	32.61	5.71
TF-IDF	33.56	5.20	35.93	6.53
Avg-DUC	34.34	6.02	37.95	7.54
SNMF	35.0	6.04	37.14	7.52
Word2Vec	34.59	5.48	36.33	6.34
PV	35.41	6.14	37.52	7.41
DocEmb	30.59	4.69	32.77	5.61
LDA	31.70	5.33	33.07	6.02
TWE	35.05	6.06	37.58	6.52
TSE	**36.28**	**6.53**	**37.96**	**7.56**
Impr	2.46	6.35	0.03	0.27

Table 2: Influence analysis of each factor for the TSE summarization, evaluated on DUC2005

Method			Rouge-1	Rouge-2	ratio 1	ratio 2
TF-IDF	sen_sim	topic				
×	√	√	35.54	6.37	2.04%	2.45%
√	×	√	34.88	5.99	3.86%	8.27%
√	√	×	35.92	6.47	0.99%	0.91%

Note that all the baselines are conducted similar with the proposed summary framework as unsupervised query-focused summarization system.

The learning rate η is set to 0.05 and gradually reduced to 0.0001 as training converge. The word2vec is additionally trained by English Gigaword Fifth Edition [3] and dimension is set to 256. The dimension of PV is set to 128, and the TWE is 64, similar as the proposed TSE model.

4.3 Experimental Results and Discussion

In this subsection, we give a report of experimental results and analysis. Table 1 shows the overall summarization performances of the proposed model and baseline models. It can be observed that our approach gives the best summary compare to any other method in ROUGE metrics over two benchmark datasets, which strongly demonstrates the outstanding performance of the proposed summarization model. Impr denotes the relative improvements over the best of the nine baselines. We can find that the proposed TSE sentence embedding consistently outperforms the baselines from 0.03% to 6.35%.

Experimental results have validated our proposed model that exploits sentence similarity and topic information can improve the overall performance. Nevertheless, they could not point out impact of the designed measure of sentence similarity. Hence, we keep consistency for our algorithm framework except for removing the part of features while calculating sentence ranking, to investigate the importance of each element as shown in Table 2. We calculate the percentage that the TSE is superior to the one neglecting one feature, denoted as ratio 1 for ROUGE-1 metrics and ratio 2 for ROUGE-2. As shown the ratio 1 is 3.86% and ratio 2 is highly up to 8.27%, it illustrates that sentence similarity computation by our proposed sentence embedding plays a consistently dominant role for the summary. On the contrary, it has improving space for utilizing topics for summary.

5 Conclusion

This work proposes a novel sentence embedding model which wisely incorporates sentence coherence and topic characteristics in the learning process. It can automatically generates distributed representations for sentences as well as assigns sentences with semantic and meaningful topics. We conduct extensive experiments on DUC query-focused summarization datasets. Utilizing the superiority of the proposed TSE that facilitates sentence ranking, the system achieves competitive performance. A promising future direction is to strengthen topic optimization during the sentence learning. With the assistance of semantic topic, we can extract sentence-based saliance topic representation as direct summary.

Acknowledgments

This work is supported by National Basic Research Program of China (973 Program, Grant No.2013CB329303), National Nature Science Foundation of China (Grant

[3] https://catalog.ldc.upenn.edu/LDC2011T07

No.61602036), and Beijing Advanced Innovation Center for Imaging Technology (BAICIT-2016007).

References

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[CLL^{+}15] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In *Proceedings of AAAI'15*, pages 2210–2216, 2015.

[CLW^{+}15] Kuan Yu Chen, Shih Hung Liu, Hsin Min Wang, Berlin Chen, and Hsin Hsi Chen. Leveraging word embeddings for spoken document summarization. *Computer Science*, 2015.

[Gal06] M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of EMNLP'07*, 2006.

[GNJ07] Surabhi Gupta, Ani Nenkova, and Dan Jurafsky. Measuring importance and query relevance in topic-focused multi-document summarization. 2007.

[HL05] Sanda Harabagiu and Finley Lacatusu. Topic themes for multi-document summarization. In *Proceedings of SIGIR'05*, pages 202–209, 2005.

[KMTD14] Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. Extractive summarization using continuous vector space models. In *Proceedings of EACL'14*, 2014.

[KNY15] Hayato Kobayashi, Masaki Noguchi, and Taichi Yatsuka. Summarization based on embedding distributions. In *Proceedings of EMNLP'15*, 2015.

[LH00] Chin Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING'00*, pages 495–501, 2000.

[LH03] Chin Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of ACL'03*, 2003.

[LLCS15] Yang Liu, Zhiyuan Liu, Tat Seng Chua, and Maosong Sun. Topical word embeddings. In *Proceedings of AAAI'15*, 2015.

[LM14] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *Computer Science*, 4:1188–1196, 2014.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013.

[MSC^{+}13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. 26:3111–3119, 2013.

[NVM06] Ani Nenkova, Lucy Vanderwende, and Kathleen Mckeown. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of SIGIR'06*, pages 573–580, 2006.

[OLLL11] You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. Applying regression models to query-focused multi-document summarization. *Information Processing & Management An International Journal*, 2011.

[PRS15] Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. Topical coherence for graph-based extractive summarization. In *Proceedings of EMNLP'15*, pages 1949–1954, 2015.

[TYC09] Jie Tang, Limin Yao, and Dewei Chen. Multi-topic based query-oriented summarization. In *Proceedings of SDM'09*, pages 1147–1158, 2009.

[WLZD08] Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of SIGIR'08*, pages 307–314. ACM, 2008.

[YCT15] Min Yang, Tianyi Cui, and Wenting Tu. Ordering-sensitive and semantic-aware topic modeling. In *Proceedings of AAAI'15*, 2015.

[YGVS07] Wen Tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. Multi-document summarization by maximizing informative content-words. In *Proceedings of IJCAI'07*, pages 1776–1782, 2007.

[YP15] Wenpeng Yin and Yulong Pei. Optimizing sentence modeling and selection for document summarization. In *Proceedings of IJCAI'15*, 2015.