Jointly Learning Topics in Sentence Embedding for Document Summarization

Yang Gao, *Member, IEEE*, Yue Xu, Heyan Huang, Qian Liu, Linjing Wei, and Luyang Liu

Abstract—Summarization systems for various applications, such as opinion mining, online news services, and answering questions, have attracted increasing attention in recent years. These tasks are complicated, and a classic representation using bag-of-words does not adequately meet the comprehensive needs of applications that rely on sentence extraction. In this paper, we focus on representing sentences as continuous vectors as a basis for measuring relevance between user needs and candidate sentences in source documents. Embedding models based on distributed vector representations are often used in the summarization community because, through cosine similarity, they simplify sentence relevance when comparing two sentences or a sentence/query and a document. However, the vector-based embedding models do not typically account for the salience of a sentence, and this is a very necessary part of document summarization. To incorporate sentence salience, we developed a model, called CCTSenEmb, that learns latent discriminative Gaussian topics in the embedding space and extended the new framework by seamlessly incorporating both topic and sentence embedding into one summarization system. To facilitate the semantic coherence between sentences in the framework of prediction-based tasks for sentence embedding, the CCTSenEmb further considers the associations between neighboring sentences. As a result, this novel sentence embedding framework combines sentence representations, word-based content, and topic assignments to predict the representation of the next sentence. A series of experiments with the DUC datasets validate CCTSenEmb's efficacy in document summarization in a query-focused extraction-based setting and an unsupervised ILP-based setting.

Index Terms—Sentence embedding, Gaussian topics, summarization, relevance, and salience

1 Introduction

In the era of the information explosion, summarization systems are playing a significant role in alleviating information overload. As such, they have been widely adopted in many applications - sentiment analysis, user profiling, online news, and question answering to name just a few. In this field, extractive summarization [1], [2] generates summaries by selecting important segments of the original input documents. Compression-based methods [3], [4] can also be used, such as deleting some uninformative or unimportant words from the selected sentences. Alternatively, abstractive summarization [5] generates new sentences and possibly re-organizes words or their orders to form fluent summaries based on a sufficient understanding of the original documents. Overall, the purpose of a text summarization system is to create a coherent, informative synopsis of the original documents. With regarding fitting user information needs, query-focused summarization is differentiated from non-query-focused or generic summarization in this area.

There are two main goals when generating a quality summary: relevance and salience. *Relevance* ensures that the summary meets the required needs of the user. *Salience* ensures that the summarized sentences capture the majority of the important information. This paper focuses on query-based extractive summarization with an emphasis on methods to improve the overall relevance and salience of the summarizations. To this end, scholars have proposed a variety of summarization models [1], [2], [6], [7], [8], [9], [10], [11] with a great deal of focus on enhancing sentence selection for salience.

Conventionally, sentence selection has relied on feature engineering to extract surface feature statistics (i.e., TF-IDF cosine similarity), which are then compared to the queries and document representations. Although this approach typically results in acceptable performance and efficiency, it does not capture the semantics of a sentence. To further improve summarization models, this research focuses on characterizing and extracting the semantics of the most representative sentences in a corpus according to specific user needs.

Recently, distributed semantic vectors in summarization systems to represent words and sentences have shown some success in selecting semantically related sentences. Using these vectors, high-dimensional and sparse linguistic text can be converted into a lower dimensional, and therefore more controllable, vector space. As a result, computing relevance based on the similarity between a specific query and the candidate sentences has become a relatively straightforward task. Motivated by the success of the Word2Vec model [12], [13], the Paragraph Vector (PV) [14] model (where the paragraph can be a sentence, a paragraph, or a document) also

- *Y. Gao, H. Huang, Q. Liu, L. Wei, and L. Liu are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail: {gyang, hhy63, weilinjing, lly_aegis}@bit.edu.cn, liuqianbit@163.com.*
- *Y. Xu is with the Faculty of Science and Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia. E-mail: yue.xu@qut.edu.au.*

Manuscript received 6 Dec. 2017; revised 29 Nov. 2018; accepted 8 Jan. 2019. Date of publication 14 Jan. 2019; date of current version 5 Mar. 2020. (Corresponding author: Yang Gao.) Recommended for acceptance by P. Cui. Digital Object Identifier no. 10.1109/TKDE.2019.2892430

Query

What hydroelectric projects are planned or in progress and what problems are associated with?

Extractive Summaries

Right Aspect:
1. Knight Piesold says the scheme involves the multi-purpose development of the Ewaso Ngiro river and includes three separate hydroelectric projects, a river transfer scheme and an irrigation project.
2. Advocates of the dam point to the reduction in flood risks, the 18bn units of electricity a year that the scheme would generate and the improved navigability of the river, enabling ocean-going ships to reach Chongqing.

Wrong Aspect:
1. The British firm, Knight Piesold and Partners, signed the contract in July 1990 with the Kenya Power and Lighting Corporation (KPLC) for consultancy services on the proposed Ewaso Ngiro hydroelectric scheme.
2. CHINA will need foreign capital and manufacturing expertise to go ahead with the Three Gorges project to control flooding in the middle reaches of the Yangtze River through construction of a huge dam, Simon Holberton reports from Hong Kong.

Reference Summaries
1. In Europe, Portugal is planning a dam at Vila Nova de Foz Coa, and Slovakia is planning a dam on the Danube River.
2. New hydroelectric projects pose social, political, economic and environmental problems.

Fig. 1. An example of the automatically-generated extractive summaries and ground truth reference summaries produced from the DUC2005 dataset with respect to the user's query. The extractive summaries on the left side, under "Right Aspect" reflect positive topics with respect to the query. The summaries on the right side, under "Wrong Aspect", show us unwanted content to the query.

contributes to predicting the next word given a sequence of words and current paragraph representation. The PV model inherits both the structure and the efficiency of the Word2Vec model, further capturing the word order to automatically generate a sentence representation. Several approaches, such as Skip-Thought vectors [15], have also been developed to map word vectors to sentence vectors through recurrent networks or convolution networks in a supervised or unsupervised manner. These methods, which are all based on sentence embedding, improve the accuracy of related summaries by directly characterizing the relevance of candidate sentences to a users query for summarization [16], [17].

However, improving the word or sentence embeddings only serves to enhance the relevance of a summary to a user's query. It does not consider the salience of the summary. The example in Fig. 1 provides more clarity. Here, the user is concerned with current and future plans for hydroelectric projects and any known problems with such projects. The ground truths, under "Reference Summaries", are a compressed version of the "Extractive Summaries" related to specific elements of the user's query. The extractive summaries are the top-ranked candidate sentences for relevance in terms of the similarity of their sentence embeddings to the embedding of the original query. These are divided into two categories: "Right Aspect" and "Wrong Aspect". The right aspect includes summarized sentences that are related to the users query. The wrong aspect includes content that is related to hydroelectric projects but does not reflect progress or problems. This example illustrates that approaches which do not consider salience in the estimation process will often retrieve somewhat related but undesirable topics given the specific needs of the user.

Hence, we intend to find solutions to enhance the salience of the selected sentences by incorporating topic analysis into a sentence embedding model. Several previous studies [18], [19], [20] have presented approaches that assign LDA-based topics to every word within the Word2Vec learning framework. But such methods do not represent sentences directly, and, therefore, cannot be seamlessly incorporated into a sentence-based summarization. To fill this gap, in this paper, we focus on introducing a new model that combines topic learning and sentence embedding into a summarization system. To model a new sentence embedding framework, a few concrete problems need to be addressed. How to construct the learning framework is the first issue, followed by which kinds of topics are appropriate to incorporate into the embedding models.

First, despite the effectiveness of the frameworks in [8], [14] in obtaining sentence embeddings, both methods primarily predict a word as an objective function and use this to estimate the minimum loss via iterations. The minimum loss is calculated in terms of a given average or by concatenating words according to context and sentence ID, and the sentence representation provides auxiliary information during the learning process. Few studies have considered the latent associations between sentences, which we argue could be very significant to sentence embeddings. For instance, sentences such as "To be or not to be" and "that is the question", often appear simultaneously, indicating that these two sentences are semantically related. Yet, both sentences rely on very different vocabularies. In cases like this, traditional embedding methods, may not identify the semantic relations between these two sentences because they rely heavily on lexical co-occurrences.

Additionally, because words with similar semantic properties are closer to each other in the embedding space, the semantic categories of words could be considered if multivariate Gaussian distributions were used to describe the central mixture of the embedding space. This fundamental assumption was recently made with a Gaussian LDA model [21]. However, the experimental results of this model in [22] showed that the generated topic embeddings were highly similar and, hence, the posterior topic proportions were almost uniform and non-discriminative. A new way to separate individual topics in the embedding space is needed.

To address the problems mentioned above, we propose a novel sentence embedding learning framework, CCTSenEmb, that enhances topic categories and sentence associations. The resulting summarizations reflect superior topical salience and sentence relevance. CCTSenEmb leverages the latent associations between sentences by directly predicting a sentence given the semantic information of the neighboring sentence. In addition, the sentence embedding space is distributed by Gaussian topics from a Gaussian mixture model (GMM), and these multiple topics are regulated by euclidean distances to ensure that these topics are discriminatory to each other. To generate the sentence embeddings, CCTSenEmb predicts the context of a sentence (i.e., the next sentence) by jointly training the content of the current sentence (i.e., the individual words in the sentence) and its discriminative topic. The main contributions of our work are:

1) We propose a novel approach to estimating sentence salience in terms of combining the importance at different levels, such as word, sentence and topic. All of the above elements comprise the salience of the summary sentences.
2) A model that considers the latent semantic associations between sentences by predicting neighborhood sentence representations. Further, Gaussian topic vectors of the sentences are thoroughly integrated into the concatenation inputs to predict the probability of the next sentence. We demonstrate that this principle is both easy to implement and effective in the sentence embedding training process.

3) The novel idea of adding a regulator into the objective function that consider the topic distance of Gaussian topics to ensure they are discriminative in the embedding space. Negative topic sampling enhance the efficacy of topic modeling based on sentence embedding.
4) Experimental results demonstrate the effectiveness of our proposed CCTSenEmb model in terms of relevance and salience in document summarization within a new integer linear programming (ILP) framework.

The rest of this paper is organized as follows. Section 2 presents the basics of word and sentence embedding models. Section 3 includes the problem formulation, followed by our proposed sentence embedding model, CCTSenEmb, in Section 4. Section 5 describes the ILP framework used for the experiments. Sections 6 and 7 report the experimental results and corresponding analysis. Finally, we conclude the paper in Section 8.

2 Background and Related Work

According to the distributional hypothesis [21], words occurring in similar contexts tend to have similar meanings. This has given rise to a central and fundamentally important technique in NLP where models learn word vectors that capture the lexical and semantic properties of the data. Word embedding, also known as a distributed representation of words, refer to the set of representation learning methods that learn the low-dimensional, real-valued dense vector representation.

2.1 Embedding Models

Word2Vec. There are two popular trends in word embedding learning methods: prediction-based [13] methods, which learn word representations by predicting the co-occurrence of words in the given context, whereas counting-based methods [23] learn word representations by factorizing the co-occurrence of words in a global matrix. Both methods learn word embeddings from term-term co-occurrence. In fact, a theoretical relation between them was discovered in [24]. Of the various word embeddings methods, our framework relies on the Word2Vec method proposed in [13]. Word2Vec is proven to be a robust baseline [24], and the log-linear models it uses to produce high-quality word embeddings are extremely computationally efficient. The approach in Word2Vec employs a sliding window that moves across the corpus. The central word is the target word; the surrounding words form the context. One model, the CBOW model, uses the average/sum of context words as the input to predict the target word. Another model, Skip-Gram, uses the target word as the input to predict each contextual word.

Paragraph Vector. [14] is an unsupervised algorithm that learns fixed-length semantic representations of variable length of text. It is a strong alternative sentence embedding model, and has been widely applied to learning representations for sequential data. Many studies [25], [26] have expanded the above models into new embedding models based on these foundations. Although a neural network is a comparable method for learning word embeddings [27], its efficiency on large datasets is an open issue. Moreover, few of these studies consider the connections between sentences while learning embeddings.

Skip-Thought. Kiros et al. [15] proposed Skip-Thought vectors, which is a method of training a sentence encoder by predicting the preceding and the following sentence. The representation learned through this objective performs competitively on a broad suite of evaluated tasks. It follows the same idea as the Skip-Gram method in the Word2Vec model but predicts the previous sentence and the next sentence based on the current sentence, rather than words. Representations for each sentence are generated by encoding the sentences with an RNN or BiRNN model. The authors of Skip-Thought assert that the vectors are sufficiently representative to capture sentences without having to learn their representations from the beginning.

2.2 Summarization Systems

Extractive summarization and abstractive summarization are two different forms of summarization systems. This paper focues on query-based extractive summarization with an emphasis on assessing the relevance of the sentences to the given query and the salience of sentences returned. Improvement in these two areas should improve the overall quality of the entire summarization. A range of features have been defined to measure relevance in multi-document summarization, including TF-IDF cosine similarity [2], [28], [29], cue words [30], topic themes [31], and concept similarity [32], among others. However, these features usually suffer from a lack of expressive capability to represent the semantics of documents. To deal with this problem, there has been a surge of recent research that focuses on using embedding models to capture linguistic regularities and semantic meaning [14], [18], [25], [33]. Neural network techniques [6], [7] have also achieved overwhelming outcomes in this area. Normally, human-generated summaries are incorporated into the framework as a learning objective schema. However, only a small portion of high-quality human labeled summaries are typically available compared to the huge amount of data that is required by CNN or RNN models. Therefore, in this paper, we have focused on unsupervised frameworks.

Embedding models [8], [16], [17], [34] for words and sentences have been used for summarization systems. DocEmb [17] formalizes the summary task as a valid submodular, objective function based on the similarities of the embeddings for unsupervised summarization. $DivSelect + CNNLM$ [8] uses a diversified selection algorithm $DivSelect$, which optimized with a convolutional network. Cao et al. [35] proposed AttSum for query-focused summarization that applies an attention mechanism to simulate the attentive reading of human behavior when a query is given. However, these methods usually reward semantic similarity without considering topic salience.

Several topic-based summarization methods have also shown their success. Parveen et al. [36] proposed an approach based on weighted graphical representations of documents generated through topic modeling. Barzilay et al. [37] use a hidden Markov model to learn the probabilistic distributional latent topics of the content (i.e., words or sentences). This and also other methods choose the “important” topics according to higher probabilities to generate summary sentences. The work proposed by Gupta et al. [38] measures topic concentration in a direct manner: a sentence is considered related to the query if it contains at least one word from the query itself.

TABLE 1
Notations

Symbol	Description
K	number of topics
W	words in the dictionary
S	sentence collection
V	dimension of vectors
T_{NG}	the negative topics of the sentence
d_i	the ith document
s	the current sentence that is predicted
s'	the previous sentence
$vec(s)$	vector of sentence s
T_s	the topic of the sentence s
$vec(T_s)$	vector of sentence topic T_s
$vec(w)$	vector of word w
π	mixture weights of GMM
μ	means of GMM
σ	covariance matrices of GMM
λ	The parameters collection of GMM
θ	the parameter for estimating sentence vector
$\mathbf{X_s}$	the context of the current sentence s, $\mathbf{X_s} = vec(T_{s'}) \oplus vec(s') \oplus vec(w_1)\oplus, \ldots, \oplus vec(w_m)$

However, this method is based on the assumption that the documents in one collection only contain only one topic. Tang et al. [39] proposed a unified probabilistic approach to uncover query-oriented topics along with four scoring methods that calculate the importance of each sentence in the document collection. Wang et al. [40] proposed a new multi-document summarization framework (SNMF) based on semantic analysis at the sentence level and symmetric non-negative matrix factorization. Symmetric matrix factorization has been shown to be equivalent to normalized spectral clustering and is used to group sentences into clusters.

Further, several approaches, such as NTM [41], TWE [18], and GMNTM [42], incorporate vector representations with topics, which combines the benefits of both semantic representation and topic classifications. By using a convolutional neural network (CNN) [43], the document categories or classifications also allow for different summary styles. To generate latent semantic topics, variational auto-encoders can be used to describe the observed sentence and its topic. Features from the model in [9] can be integrated into a new framework to estimate salience for extraction-based summarization.

The above studies motivated our investigation into an elegant way to combine query-related topics, sentence relevance, and salience within one summarization system. In addition to incorporating topics, our novel sentence embedding model highlights the associations between sentences, which gives the nearest neighbors relatively similar sentence vector representations. Using this new sentence embedding technique, we expect the generated summaries to be more salient and relevant than conventional sentence embedding models.

3 Problem Formulation

This section introduces the key concepts involved in the CCTSenEmb model. Table 1 lists the notations used in this paper.

Consider a document collection $D = \{d_1, d_2, \ldots, d_M\}$ that contains a set of sentences S and each $s \in S$. The collection includes K number of topics; hence, we assume that the sentences are distributed as Gaussian topics over the space. To enhance the semantic association between sentences, which is regarded as coherence in this paper, the proposed CCTSenEmb alternatively predicts sentences for obtaining sentence representations during the learning process instead of using prediction task based on words as the Word2Vec or the PV model in the Section 2. In each iteration, the target sentence is not only predicted by the words in the previous sentence, the sentence itself, and its vectorized topics, but also by sampled sentences chosen from negative topics. These concepts are illustrated in the following definitions.

Definition 1 (Context). *Given a previous sentence s' and a current sentence s, each serves as the neighboring context for the other. Context reflects hidden associations between neighboring sentences.*

Definition 2 (Content). *The content of sentence s contains the sequence of the embeddings of its words $vec(w_1), \ldots, vec(w_m)$.*

Definition 3 (Positive Topic). *Let T_s denote the topic assigned to sentence s; T_s is the positive topic for sentence s.*

Definition 4 (Negative Topic). *Once the positive T_s has been defined, the other topics in the number of K topics of the collection become the negative topics for sentence s, denoted as T_{NG}.*

Definition 5 (Label). *Given sentence s, s^+ is a set of sentences for s that have been assigned a positive topic T_s. The set of sentences for sentence s that have been assigned as negative topics T_{NG} is denoted as s^-. The label of sentence $\widetilde{s} \in S$ is defined as*

$$l(\widetilde{s}) = \begin{cases} 1, \widetilde{s} \in s^+; \\ 0, \widetilde{s} \in s^-; \end{cases}. \tag{1}$$

4 The Proposed CCTSenEmb Model

As mentioned in the introduction, neighboring sentences often contain very different vocabularies even though they may be relatively semantically close. Normal embedding methods are unable to identify the semantic similarities between these sentences, and, therefore, the resulting sentence representations can be quite far away from each other in the vector space.

To solve this problem, the CCTSenEmb model identifies and reflects the associations between neighboring sentences during the sentence prediction process through a novel embedding method that considers words, sentences, and sentence topics. Neighboring sentences are predicted by exploiting sentence-to-sentence latent associations. The Gaussian topic distributions of the sentences are integrated into the input vectors and, as the sentence vectors update, the discriminative topic distributions are iteratively optimized during the learning process. Vectors for words, sentences, and positive vectorized topics are all incorporated as semantic content to make positive predictions, while the sampled sentences from other topics are used to make negative predictions.

The following subsections present the details of the CCTSenEmb model. The models overall architecture for learning sentence representations is shown in Fig. 2.

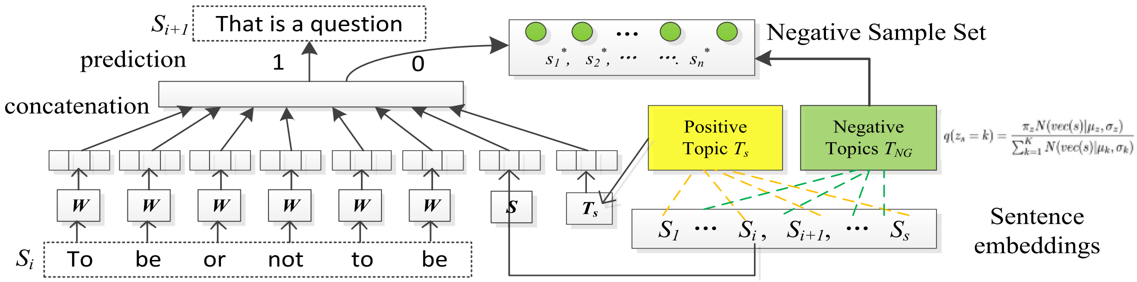


Fig. 2. The structure of the proposed CCTSenEmb model.

4.1 Discriminative Gaussian Topic

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

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

represents the probability distribution for sampling a vector x from the GMM.

Generally speaking, the GMM builds the objective function by estimating the maximum likelihood:

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} logp(x|\theta), \tag{3}$$

where θ represents the parameter of GMM, m denotes the number of the sample data, and $p(\cdot)$ is the likelihood function.

As mentioned in the introduction, traditional GMM topic embedding methods suffer from highly similar generated topic embeddings in the vector space. To address the problem of generating GMM topics that are possibly overlapping to each other, topic distances are incorporated in terms of L2-norm into the objective function to generate discriminative topics.

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} logp(x|\theta) + \sum_{i}\sum_{j} ||\mu_i - \mu_j||_2^2. \tag{4}$$

Given a GMM $G = \{N(\mu_i, \sigma_i)|i = 1, 2, 3, \ldots K\}$, the log likelihood function of a specific Gaussian distribution $N(\mu_k, \sigma_k)$ in the GMM is denoted as:

$$\begin{aligned} LL &= log\prod_{i=1}^{N_k} \frac{1}{\sqrt{[(2\pi)^d|\sigma_k|]}} \exp\Bigg\{ -\frac{1}{2}\gamma(i,k)(x_i-\mu_k)^\top \sigma_k^{-1} \\ &\quad (x_i - \mu_k)\Bigg\} \\ &= -\frac{N_k}{2}\cdot \ln[(2\pi)^d \cdot |\sigma_k|] - \frac{1}{2}\cdot \sum_{i=1}^{n} \gamma(i,k)(x_i-\mu_k)^\top \\ &\quad \sigma_k^{-1}(x_i-\mu_k). \end{aligned} \tag{5}$$

The added regularization term maximizes the euclidean distances between the Gaussian topics. Therefore, the new objective function becomes

$$\mathcal{L}(\theta) = LL + \frac{\alpha}{2}\sum_{j=1}^{m} ||\mu_k - \mu_j||^2, \tag{6}$$

where α is co-efficient of the regularization term.

To maximize $\mathcal{L}$, a derivation is taken with respect to μ and σ.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_k} &= -\frac{1}{2}\sum_{i=1}^{N_k} \gamma(i,k)[\sigma_k^{-1} + (\sigma_k^{-1})^\top](x_i - \mu_k) \\ &\quad + \alpha \sum_{j=1}^{K}(\mu_k - \mu_j) \end{aligned} \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_k} = \frac{\partial LL}{\partial \sigma_k}. \tag{8}$$

Let $\frac{\partial \mathcal{L}}{\partial \mu_k} = 0$ and $\frac{\partial \mathcal{L}}{\partial \sigma_k} = 0$, and we have solutions for μ_k and σ_k.

$$\begin{aligned} \mu_k = &\left[2\alpha K\cdot \mathbf{I} - \left[\sigma_k^{-1} + (\sigma_k^{-1})^\top\right]\left[\sum_{i=1}^{N_k}\gamma(i,k)\right]\right]^{-1} \cdot \\ &\left[2\alpha\sum_{j=1}^{K}\mu_j - \left[\sigma_k^{-1} + (\sigma_k^{-1})^\top\right]\cdot\left[\sum_{i=1}^{N_k}\gamma(i,k)x_i\right]\right] \end{aligned} \tag{9}$$

$$\sigma_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\gamma(i,k)(x_i-\mu_k)^\top(x_i-\mu_k). \tag{10}$$

Consequently, an expectation maximization (EM) algorithm can be leveraged to find the optimal solution. The posterior distribution of the topics in each sentence s can be inferred by

$$q(z_s = k) = \frac{\pi_z N(vec(s)|\mu_z, \sigma_z)}{\sum_{k=1}^{K} N(vec(s)|\mu_k, \sigma_k)}. \tag{11}$$

Based on the distribution, the max probability index is selected as the positive topic of sentence s. The topic maximizing Equation (11) for sentence s is the positive topic T_s; the other topics are negative topics for s. However, optimizing these objectives is non-trivial and very time-consuming. To reduce the learning complexity, we employ the idea of negative sampling [12], [13].

4.2 Negative Topic-Based Sampling

The idea of negative sampling is to approximate the costly probabilities of softmax with some simple negative samples. Learning can then be conducted by optimizing a point-wise classification loss. For a center vertex s , high-quality negatives should be sentences that are dissimilar to s. Some heuristics have been applied to achieve this goal, such as sampling from probability-based non-uniform distributions [13]. Here, we propose a more adaptive sampling method that caters to the dissimilarities among different topics for sentence embedding.

As Fig. 2 shows, sentences across the entire collection can be divided into positive topics and negative topics, regarding to the current sentence s. Specifically, if s is assigned with a specific topic T_s, then let T_s be the positive topic of sentence s. All sentences with a topic assignment similar to s belong to the positive topic collection. Similarly, let T_{NG} be the set of negative topics in regards to the sentence s, which equals $T \backslash T_s$. All the sentences with a topic assignment dissimilar to s belong to the negative topic collection.

Given a center sentence s, the negative samples from the set of sentences with negative topics are randomly chosen. Using this method yields high-quality and diverse negative samples since the discriminative Gaussian topics operation guarantees that dissimilar sentences are located in different topics in a probabilistic way. Specifically, the negative topics that are randomly chosen from the set of T_{NG} are denoted as $\mathbb{N}(T^* \sim T_{NG})$. A negative sentence is then sampled from this collection. Thus, the complete process of negative topic sampling for the sentence follows $\mathbb{E}(s^* \sim \mathbb{N}(T^* \sim T_{NG}))$. The details of approximating the conditional probability are then defined in Equation (13).

4.3 Sentence Prediction

The assumption within CCTSenEmb is that sentences are coherent and associated with their neighbors. Consequently, a sentence is modeled as a prediction task based on the semantic information of the previous sentences. The sentence is then represented as a combination of the sentence's topic, its representation, and its content.

Let $\mathbf{X_s}$ be the context vector of the previous sentence to the current sentence s. The vector is denoted as a concatenation of the above information in the previous sentence, where $\mathbf{X_s} = vec(T_{s'}) \oplus vec(s') \oplus vec(w_1) \oplus, \ldots, \oplus vec(w_m)$. The semantic vectors, including the sentence topic vector, the sentence vector, and the word vector of the content are all incorporated. In our experiments, we empirically set the sentence to a fixed number of 20 words. Words in excess of 20 in long sentences were severed, and zeros were used as placeholders in sentences of less than 20 words.

A description of the procedure to generate the embeddings for words, sentences, and topics follows. Given the GMM λ, the generative process is first initialized as follows: for each sentence s in S, its previous sentence is s', and its topic $T_{s'}$ is chosen from the multinomial distribution $\pi := (\pi_1, \pi_2, \ldots, \pi_T)$. Its vector representation $vec(s')$ is sampled from $N(\mu_{T_{s'}}, \sigma_{T_{s'}})$. Each sentence is then predicted by $sigmoid(\theta \mathbf{X_s}^\top)$.

Given these Gaussian topics, the next prediction process concentrates on exploiting the latent relationship between a sentence and its context in the collection. Subsequently, the probability of the current sentence s (i.e., the target sentence) is predicted by the information from the previous sentence. Therefore, the overall objective of CCTSenEmb is to maximize this probability

$$G = \prod_{s \in S} g(s) = \prod_{s \in S} \prod_{u \in \{s \cup s^-\}} p(u|\mathbf{X_s}). \tag{12}$$

Rather than using a softmax function for the prediction probability, we directly implement the negative sampling approximation. In the above equation, S is the collection of all sentences, and s is the current sentence that is correctly predicted, and s^- is the set of negative samples selected by $\mathbb{E}(s^* \sim \mathbb{N}(T^* \sim T_{NG}))$ as introduced in the previous Section 4.2. The prediction objective function of sentence s is $g(s) = \prod_{s \in S} p(u|\mathbf{X_s})$, and the probability function is defined as

$$p(u|\mathbf{X_s}) = \begin{cases} sigmoid(\theta^u \mathbf{X_s}^\top), l(\widetilde{u}) = 1 \\ 1 - sigmoid(\theta^u \mathbf{X_s}^\top), l(\widetilde{u}) = 0 \end{cases} \tag{13}$$

or, in a unified equation, this is written as

$$p(u|X_s) = [sigmoid(\theta^u \mathbf{X_s}^\top)]^{l(\widetilde{u})} \cdot [1 - sigmoid(\theta^u \mathbf{X_s}^\top)]^{1-l(\widetilde{u})}, \tag{14}$$

where $sigmoid(z) = 1/(1+\exp(-z))$ and $\theta^u \in R^V$ is the parameter of sentence s.

The log-likelihood of the objective function is then derived, defined as

$$\begin{aligned} \mathcal{L} = \sum_{s \in S} &\{ l(\widetilde{s}) \log [sigmoid(\theta^s \mathbf{X_s}^\top)] \\ &+ \sum_{u \in s^-} (1 - l(\widetilde{u}))(n\mathbb{E}(s^* \sim \mathbb{N}(T^* \sim T_{NG}))) \\ &\log [1 - sigmoid(\theta^u \mathbf{X_s}^\top)] \}, \end{aligned} \tag{15}$$

where $n\mathbb{E}(\cdot)$ is the number of n negative samples as in Definition 1, and n is empirically set to 10.

4.4 Parameter Estimation

Given the model parameters and the vectors, we can infer the posterior probability distribution of topics as Equation (11). Thus, the topic for sentences are obtained after the initialization, and all the vectors are ready for iterative update as outlined in the following two stages. Stage I, the likelihood of the model is maximized with respect to $\lambda = \{\pi_k, \mu_k, \sigma_k\}$. Since this characterizes a GMM, the procedure can be implemented with the EM algorithm. In Stage II, the model likelihood is maximized with respect to $\theta^u, \mathbf{X_s}$. This procedure can be implemented with stochastic gradient descent (SGD). Stage I and Stage II are alternatively executed in an alternating fashion until the parameters converge. The algorithm is summarized in Algorithm 1.

SGD is used to update the parameters of the GMM, $\{\theta^u, \mathbf{X_s}\}$. Given λ, the gradient of θ^u is calculated using back propagation based on the objective in Equation (15) as (and similar for $\mathbf{X_s}$)

$$\frac{\partial \mathcal{L}(\mathbf{X_s}, \theta^u)}{\partial \theta^u} = [l(\widetilde{u}) - sigmoid(\theta^u \mathbf{X_s}^\top)] \mathbf{X_s}$$

$$\frac{\partial \mathcal{L}(\mathbf{X_s}, \theta^u)}{\partial \mathbf{X_s}} = [l(\widetilde{u}) - sigmoid(\theta^u \mathbf{X_s}^\top)] \theta^u \tag{16}$$

and the vector of the sentence $vec(s)$ is finally extracted from $\mathbf{X_s}$.

Algorithm 1. CCTSenEmb Model

Input:
Sentences S, $|W|$ words contained in dictionary W, the learning rate η, the dimension of the vector V and the number of topics K.

Output:
Topic representations $vec(T_s)$, vector representations $vec(w)$ and $vec(s)$.

1: Randomly initialize parameters (i.e., λ, θ^u) and vectors (i.e., topics, words and sentences)
2: Repeat until converge

- Fix parameters θ^u and $\mathbf{X_s}$, 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:
  $N_k = \sum_{i=1}^{N}\gamma(i,k)$;
  Compute μ_k and σ_k by Equation (9) and (10);
  $\pi_k = \frac{N_k}{N}$
- Fixing parameters λ, run the SGD algorithm:
  Update θ, X_s by:
  $l_1 = sigmoid(\theta^u \mathbf{X_s}^\top)$; $l_2 = \eta(l(\widetilde{u}) - l_1)$;
  $\mathbf{X_s} := \mathbf{X_s} + l_2\theta^u$; $\theta^u := \theta^u + l_2\mathbf{X_s}$

5 ILP-Based Multi-Document Summarization

The relevance of a sentence to a query is primarily measured with vector-based cosine similarity [16], which is a promising measure for computing query-sentence relatedness in summarization tasks. Additionally, statistical features (i.e., TF-IDF scores [28]) and prior knowledge (i.e., sentence position [44]) are jointly collected. Hence, the salience and relevance of a sentence j, noted as Sal_j, is initially calculated by

$$Sal_j = p\left(\sum_k I(w_k) + R(s_j,q)\right), \tag{17}$$

where q is the user-specific query and w_k is the word in sentence j. The relevance $R(s_j,q)$ is the similarity of the sentence j and the query q, which is calculated by cosine similarity. $I(w_k)$ is the importance of the word k, such as TF-IDF(w_k). Of particular note is that positional information is used to generate

$$p = \begin{cases} C^{SID}, & ifSID < \overline{SID} \\ C^{\overline{SID}}, & otherwise \end{cases} \tag{18}$$

where SID is the sentence ID in each document starting from 0, and C is 0.6 in this paper. $\overline{SID}$ is a constant position in the sentence, defined as 10 in this paper.

In particular, topic salience t_i is also considered in the summarization. t_i is estimated via the topic frequency discovered by the CCTSenEmb model across the whole collection, where $t_i = \frac{n_i}{|S|}$, n_i is the number of sentences assigned with topic i and $|S|$ represents the number of all the sentences in the collection. The overall objective function of this optimization formulation is to incorporate that salience of sentences and topics, and sentence relevance, while removing redundancy in the summary. This process can be formulated as an ILP problem [45]:

$$max\left\{\sum_i \alpha_i t_i + \sum_j \beta_j Sal_j - \sum_{j<j'} \beta_{jj'} R_{s_j s_{j'}}\right\}, \tag{19}$$

where α_i and β_j are selection indicators for topic i and sentence j, respectively. t_i and Sal_j are the salience scores for the topics and sentences. $\beta_{jj'}$ is the co-occurrence indicator of the pair $(s_j, s_{j'})$. $R_{s_j s_{j'}}$ is the similarity of the sentence pair $(s_j, s_{j'})$. The similarity is calculated through pairwise cosine similarity. This objective maximizes the salience score of the selected topics and related sentences but penalizes the selection of similar sentence pairs.

$$\beta_j Asso_{ij} \leqslant \alpha_i \tag{20}$$

$$\sum_j \beta_j Asso_{ij} \geqslant \alpha_i \tag{21}$$

$$\beta_{jj'} - \beta_j \leqslant 0; \beta_{jj'} - \beta_j' \leqslant 0; \tag{22}$$

$$\beta_j + \beta_j' - \beta_{jj'} \leqslant 1 \tag{23}$$

$$\beta_j \leqslant 0 \quad if \quad sim(s_j,q) \geqslant 0.5 \tag{24}$$

$$\sum_j l_j\beta_j \leqslant L \tag{25}$$

$$\{\alpha_i, \beta_j, \beta_{jj'}\} \in \{0,1\}, \quad \forall i,j,j < j'. \tag{26}$$

Inequalities Equations (20), (21) associate the sentences and topics, where $Asso_{ij}$ indicates that the topic i belongs to sentence j. This ensures that selecting a sentence leads to the selection of the topic it belongs to, and selecting a topic only happens when it is present in at least one of the selected sentences. Some constraints have been added to the ILP framework, such as Equations (22), (23) to control the co-occurrence relation between two sentences. Constraint Equation (22) states that if the sentence s_j and $s_{j'}$ co-occur in the summary (i.e., $\beta_{jj'}$ = 1), then they must be included them individually (i.e., β_j = 1 and β_j' = 1). Constraint Equation (23) is the inverse of Equation (22). Inequality Equation (24) prevents similar sentences from being selected (cosine similarity > 0.5), and L limits the length of the generated summary by Equation (25).

The objective function and constraints are linear; therefore, the optimization can be solved by ILP solvers. In this paper, we used the solvers implemented by Gurobi.[1]

6 Experimental Setup

This section presents the experiments conducted to evaluate the performance of the CCTSemEmb on a query-focused multi-document summarization task. Three hypotheses (H1, H2, and H3) were formulated to test the effectiveness of the framework:

1) *H1*: Incorporating topic salience and sentence relevance into an ILP-based summarization method that

1. https://www.gurobi.com/

uses the CCTSenEmb model will perform comparatively better than the baseline models;

2) *H2*: Incorporating discriminative topics into the embedding model and using a negative topic sampling strategy are effective for obtaining quality sentence embeddings;
3) *H3*: CCTSenEmb will generate more accurate sentence representations than the state-of-the-art sentence embedding models and ensure the most relevant sentences are selected.

Note that the baselines and the proposed summary framework were all evaluated within an *unsupervised*, query-focused summarization system.

6.1 Datasets

The DUC2005 and 2006 datasets[2] were used for query-focused multi-document summarization. The documents were drawn from the news domain and grouped into thematic clusters. Both datasets contain 50 query-focused summarization tasks. A relevant document cluster comprising 25-50 documents was assumed to be "retrieved" for each query from the DUC2005 dataset. Each of the document clusters within the DUC2006 dataset contained 25 documents. The task was to generate a summary from the document cluster to answer the query.[3] Each cluster also contained several articles written by various authors, which are served as the ground truth for evaluation. The resulting summaries were limited to 250 words.

6.2 Evaluation Metrics

ROUGE [46] metrics were used to evaluate the experimental results. ROUGE-N is an n-gram recall measure that is used to evaluate the quality of a summarization by counting the number of overlapping units, such as n-grams. ROUGE-1 emphasizes the co-occurrence of words in the candidate and reference summaries. ROUGE-2 focuses more on the readability of the candidate summary. It is defined as follows:

$$Rouge - N = \frac{\sum_{S \in Ref} \sum_{gram_n \in S} CountMatch(gram_n)}{\sum_{S \in Ref} \sum_{gram_n \in S} Match(gram_n)}, \quad (27)$$

where n stands for the length of the n-gram, Ref is the set of reference summaries. $CountMatch(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, and $Count(gram_n)$ is the number of n-grams in the reference summaries.

6.3 Baseline Models

The baseline models fall into three categories. The first category included conventional summarization methods:

Lead selects the leading sentences to form a summary and is DUC's official baseline. *TF-IDF* [28] ranks sentences according to their TF-IDF cosine similarity to the query. *SNMF* [40] uses non-negative matrix factorization (SNMF) to cluster sentences and then selects multi-coverage summary sentences. *MultiMR* [29] is a graph-based manifold ranking method that leverages sentence-to-sentence and sentence-to-query relationships. *DocEmb* [17] ranks sentences according to deep neural network features. It includes a valid submodular objective function based on embedding distributions. We used a linear scaling function, similar to [17], to calculate the KL-divergence of the document distance.

The ILP-based CCTSenEmbs sentence selection algorithm was implemented in ILP. The similarity $R(.)$ was generated by *CCTSenEmb*, then compared with the following word embedding and sentence embedding baselines.

Word2Vec and the *PV* model are introduced in [12], [14]. The settings of both models are introduced in the following subsections. *Skip-Thought* [15] learns sentence embeddings by predicting the previous and next sentences based on the current sentence. The Skip-Thought model was pre-trained on a large collection of novels called BookCorpus,[4] then incrementally trained to represent the sentences in the summarization collections from the DUC2005 and DUC2006 datasets. *TWE* [18] employs latent topic models to assign topics for each word in the text corpus and learns topical word embeddings based on both words and their topics. *CCT_RandomSampling(CCT_RS)* replaces the negative topic sampling strategy in the CCTSenEmb with a random sample of all noisy sentences as a counterpart baseline.

6.4 Parameter Settings

CCTSenEmbs learning rate was set to 0.05 and gradually reduced to 0.0001 as the training converged. The CCTSenEmb, PV, and TWE models were trained on DUC 2005 and DUC2006. Word2Vec was additionally trained on the English Gigaword Fifth Edition[5] dataset, which consists of a standard Gigaword that contains around 9.9 million news articles sourced from various domestic and international sources. Using this dataset, we trained 256-dimensional vectors for 1.53 million words. We also implemented our negative sampling method and the Skip-Gram architecture to generate the word embeddings. Dimensionality in the PV model was set to 128, and to 64 for the TWE and the CCTSenEmb models. The number of topics in CCTSenEmb and TWE were similar to their vector dimensions, i.e., 64. The word embedding representations were also retrained in both TWE and the CCTSenEmb on a collection of DUC datasets with 64 dimension.

7 Experimental Results and Discussion

This section presents the comparison between CCTSenEmb and the baselines for multi-document summarization and includes an analysis the performance of the proposed model.

7.1 Unsupervised Query-Focused Document Summarization

Table 2 shows the overall unsupervised summarization performance of CCTSenEmb and the baseline models. According to the ROUGE metrics on both benchmark datasets, CCTSenEmb generated the best summaries of all the methods, strongly demonstrating its outstanding performance. $Improvement\%$ represents a relative improvement over the best of the nine baselines. We can see that the proposed

2. http://duc.nist.gov/data.html
3. A query is called a "narrative" in the DUC datasets. For other kind of query-focused summarization datasets, the format of queries can be varied.
4. https://github.com/ryankiros/skip-thoughts
5. https://catalog.ldc.upenn.edu/LDC2011T07

TABLE 2
Overall ROUGE Evaluation (%) of the Different Models on the DUC2005 and DUC2006 Datasets

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
SNMF	35.0	6.04	37.14	7.52
MultiMR	35.58	6.81	38.57	7.75
DocEmb	34.34	6.02	37.95	7.24
ILP+Word2Vec	36.63	6.48	37.91	7.03
ILP+PV	37.1	6.95	38.32	7.82
ILP+Skip-thought	37.62	6.93	38.77	7.41
ILP+TWE	35.9	6.49	38.07	7.01
ILP+CCTSenEmb	**38.65**	**7.73**	**39.61**	**8.43**
Improvement%	2.74	11.2	2.27	7.80

ILP+CCTSenEmb model consistently outperformed the best of the baselines, ranging from 2.27 to 11.2 percent. Overall, the statistics validate the first hypothesis, H1.

The specific improvement with respect to ROUGE-2 was much better than the improvement for ROUGE-1, which indicates that CCTSenEmb is able to enhance the readability of the summaries to some extent. In evaluating the utility of four state-of-the-art embedding methods working in conjunction with the cosine similarity measure for summarization, all four (i.e., Word2Vec with Skip-Gram, PV, Skip-Thought, and TWE) produced comparable results. The TWE model assigns different topics to individual words; however, is was difficult to implement topic representation while selecting summary sentences. Therefore, only word embeddings were used with the ILP+TWE model. ILP+Skip-Thought performed the best in terms of the ROUGE-1 metric, and ILP+PV performed the best in ROUGE-2 on both datasets. This result demonstrates that using sentence embedding as a representation for summarization achieves better results than using word embeddings

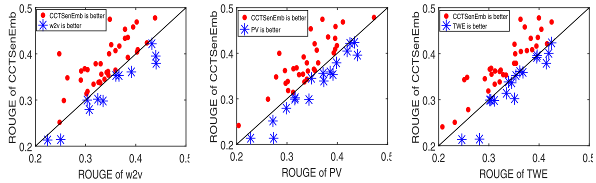
The results in Table 2 demonstrate that the summarization systems based on an ILP construction consistently outperformed the classical baselines, including the graph-based method (MultiMR), the feature-based method (TF-IDF), and the submodular model (DocEmb). These statistics further emphasize the first hypothesis.

7.2 The Quality of Sentence Embedding

To individually evaluate the effectiveness of CCTSenEmb's sentence embedding in unsupervised query-focused summarization tasks, we neglected the topic salience estimation t_i in Equation (17) and only used the salience Sal_j parameter. Sentence selection was implemented in ILP. The results appear in Figs. 3, 4, and in Table 4.

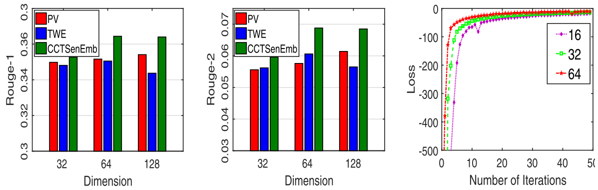
Fig. 3 shows that the CCTSenEmb model (denoted by red circles) delivered outstanding performance in the majority of cases compared to Word2Vec, PV, and TWE (denoted by blue stars). The relative improvements are obvious. This analysis supports the third hypothesis.

The sentence embedding models, i.e., ILP+PV, ILP+Skip-Thought and the proposed CCTSenEmb consistently outperformed the word embedding-based models, i.e., DocEmb, ILP+Word2Vec, and ILP+TWE, in Table 2. As we can see,



(a) CCTSenEmb v.s. Word2Vec (b) CCTSenEmb v.s. PV (c) CCTSenEmb v.s. TWE

Fig. 3. The ROUGE-1 evaluation of CCTSenEmb and the state-of-the-art embedding models on each cluster of DUC2005. # Cluster = 50.



(a) ROUGE-1 (b) ROUGE-2 (c) Iterations of loss

Fig. 4. The results of each embedding model trained in 32, 64, and 128 dimensions. (a) Results in terms of ROUGE-1. (b) Results in terms of ROUGE-2. (c) The value of the loss function in iterations of CCTSenEmb.

word salience with all these methods was similar; the only difference was their relevance to the query computed by sentence embedding-based similarity. These results demonstrate that sentence embeddings (i.e., PV, Skip-Thought, and CCTSenEmb) are more appropriate for computing sentence relevance for summaries than word embeddings (i.e., Word2-Vec, TWE). These results also show that the quality of embeddings directly influences the summarization performance in terms of relevance computing between sentences. Further, using word or sentence embeddings to capture semantics improves the quality of summaries over more traditional approaches (i.e., TF-IDF and LDA).

7.3 Dimension of Embedding Models

CCTSenEmb also outperformed the other models, such as PV and TWE, across all dimensions according to the results shown in Fig. 4. In the Fig. 4c shows that the proposed model was the fastest to converge in 64 dimensions than in other circumstances. From the perspective of topics, both TWE and CCTSenEmb reached their peak at 64 dimensions, while the optimal dimension for PV was 128. This indicates that the proposed CCTSenEmb generates denser vectors than the PV, which can accelerate task-oriented uses as well as enhance summarization accuracy. The varying performance of these models in different dimensionalities suggests that topic categories can represent data in an abstract manner and, therefore, may have the potential to provide a high level of guidance for vector embedding. In effect, the PV model

TABLE 3
ROUGE Evaluation (%) of the Topic Sampling Methods on the DUC2005 and DUC2006 Datasets

Method	DUC2005		DUC2006	
	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
CCT_RS_no_topic	36.44	6.88	38.46	7.26
CCTSenEmb_no_topic	37.01	6.98	38.66	7.83
CCTSenEmb+Gtopic	37.28	7.03	38.96	7.56
CCTSenEmb+Dtopic	**38.65**	**7.73**	**39.61**	**8.43**

TABLE 4
The effect of Sampling Different Numbers of Negative Topics (Removing Topic Salience t_i) on the DUC2005 Dataset

#. Neg.Topics	ROUGE-1	ROUGE-2
3	34.86	5.88
6	34.92	5.97
8	36.10	6.48
10	37.01	6.98
15	36.59	6.47

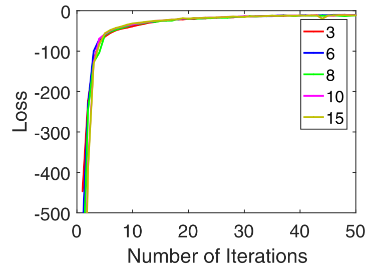


Fig. 5. Losses with different numbers of negative topic samples.

can be explained as a special case of topic embedding, where each sentence contains a unique topic. However, it is worth noting that PV was less efficient than the CCTSemEmb. These results support the third hypothesis.

7.4 Negative Topic Sampling

In Table 3, both CCT_RS_no_topic and CCTSenEmb_no_-topic only used sentence salience and relevance of Sal_j and remove redundancy in the ILP framework, rather than using topic salience t_i in Equation (19). The CCTSenEmb_no_topic, which relies on negative topic sampling, performed comparatively better than the CCT_RS_no_topic method by between 2.46 to 10.70 percent. These statistics demonstrate that the strategy of negative topic-based sampling improves the quality of the sentence embedding. Gaussian topics separate positive topics from negative ones, which provides guidance in choosing negative samples instead of using random operations, which accounts for the improvement. Adding an ILP strategy to the topic salience estimation greatly enhanced the summary results in terms of ROUGE-1 and ROUGE-2 for both datasets. The results for CCTSenEmb+Dtopic in Table 3, which includes the proposed *D*iscriminative topics, shows an obvious improvement over the baselines as a result of including topic salience in the summarization. These results strongly support the second hypothesis, H2.

Additionally, we compare CCTSenEmb+Gtopic, which only includes *G*aussian topics without implementing a distance regularizer, to the two baselines along with the proposed CCTSenEmb+Dtopic. The results demonstrate that the traditional Gaussian topics give little improvement to summarization performance. CCTSenEmb+Dtopic was the only method to make a significant improvement because it incorporates a discriminative regularizer.

Table 4 lists the results with a varying number of negative topic samples. The results demonstrate that a certain number of negative samples positively affects sentence embedding. As the number of negative samples rises to a point, the quality of sentence embedding consequently improves, as does summarization performance. However, at 15 topics, the quality fell. This indicates that, while it is better to use negative samples from a wider range of negative topics, too many negative samples generate noise in the estimation process, and that affects the probability of the target sentences. Introducing noise is likely in the following case: a specific sentence may belong to multiple topics due to the properties of the GMM model, which, could simultaneously be distributed in the positive and negative topics. As the number of negative samples increases, it becomes more likely that a positive sentence will be sampled from the negative topics and labeled as a negative sample. As a result, it is a fault during the learning process. Therefore, the proper number of negative samples is needed.

The results of the loss function tests, shown in Fig. 5, also show that CCTSenEmbs convergence speed is not very sensitive to the number of sampled negative topics. That is to say, the performance of CCTSenEmb is relatively stable and robust to the parameter settings for learning sentence embeddings. This result further validates the effectiveness of the proposed negative topic sampling approach in support of the second hypothesis.

7.5 Topic Inspection and Visualization

The case study in Fig. 6 shows some of the sample sentences for each topic, which help to intuitively understand the latent topics. From these samples, we observed that the derived topics have semantic and converged meanings. More importantly, sentence-based topic representation is able to dramatically enrich topic interpretation. For instance, the topic, “economic problem” is not only relevant to economics but also to its problems. Enriching topic interpretation precisely addresses the objective mentioned in the introduction facilitating a complex topic matching calculation for summarization. Similar benefits were mentioned in [47], who argue that topics can be better represented by sentence summaries than by comparing individual words.

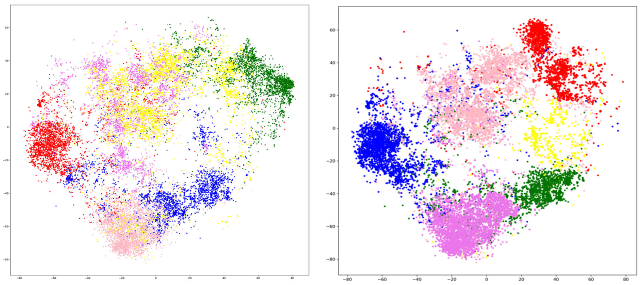
Fig. 7 visualizes the sentence representations that were projected onto two-dimensional space with TSNE toolkit.[6] Figs 7a and 7b represents topic distributions generated by CCTSenEmb+Gtopic and CCTSenEmb+Dtopic, respectively.[7] From Fig. 7a, we can see that topics generated by CCTSenEmb+Gtopic are sparsely distributed, many of which are overlapping. In contrast, in Fig. 7b, the topics from our proposed CCTSenEmb+Dtopic are separately distributed, and many of the overlaps are eliminated. We can conclude that the generated topics using CCTSenEmb+Dtopic are more discriminative when representing sentences from the whole collection. Therefore, sentences from the same topic are semantically closer to each other, and otherwise dissimilar from different topics - a beneficial property when computing most relevant sentences to the provided query.

6. http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

7. Noted that the topics in (a) and (b) could not be identically similar since were trained differently. Hence, we randomly chose six topics from CCTSenEmb+Gtopic to visualize. Then, we found as many of the same sentences as possible from CCTSenEmb+Dtopic and visualized those with their corresponding new topics.

Topic- Food Calories	Topic-Economic Problem	Topic-Wildlife Rescue
1. *Under the regulations, the new* ***food*** *labels would display prominently the total* ***calories*** *from fat, and the amounts of* ***dietary*** *fiber, cholesterol and saturated fat contained in the* ***product****.*	1. *These omissions undermine Oxfam's assertion that the structural adjustment policies urged on Africa by the* ***World Bank*** *and the* ***IMF*** *have not halted the continent's decline.*	1. *It takes an unconventional man to rescue a nation's* ***wildlife*** *from the clutches of ivory poachers, smuggling syndicates and real estate barons.*
2. *The* ***oyster mushroom****, which she says tastes a little like* ***bacon*** *when sauteed on medium high heat for a few minutes, remains one of her favorites.*	2. *The* ***IMF's*** *focus is on shorter-term macro-economic* ***problems****, while the* ***World Bank****'s is on longer-term structural* ***difficulties****.*	2. *A typical annual* ***wildlife*** *quota per district might include a variety of* ***wildlife****, such as seven* ***elephants****, 20* ***buffalo*** *and 20* ***baboons****.*
3. ***Sugar Free*** *Fun* ***cakes*** *are priced from 35; one-half gallon of homemade* ***ice cream*** *is 15.*	3. *The criticism,which is not new of* ***World Bank****-sponsored economic adjustment programmes, is supported by the facts*	3. ***Wildlife*** *and fisheries secretary Joe Herring estimated a 50 per cent decline in* ***alligator*** *industry.*

Fig. 6. Topic interpretation by sentences.



(a) (b)

Fig. 7. Visualization of topic distribution in vector space. Each dot in the figure represents one sentence, and each color represents one topic. Sentences in the same topic are represented by the same color. (a) is the topic distribution generated by CCTSenEmb+Gtopic, and (b) is the topic distribution generated by the proposed CCTSenEmb+Dtopic.

8 Conclusion

This paper introduces CCTSenEmb - a sentence embedding framework that jointly learns topics, sentences, and words in a unified document summarization system with a negative topic sampling strategy for estimation. Overall, the CCTSenEmb meets our initial goal of discovering the hidden associations between sentences and integrating discriminative topics into the learning process. Further, this novel approach shows a significant improvement over various baselines tested with DUC datasets. CCTSenEmb is helpful for detecting topics and extracting coherent, relevant summaries. Although this particular study mainly focuses on summarization, the methods presented are broadly applicable to other content-based applications, such as information retrieval or answering questions.

Our future research will concentrate on upgrading the topic learning process to broaden CCTSenEmbs usability. Additionally, studying Gaussian topic-based sentence embedding revealed some interesting issues with the efficiency of estimating Gaussian topics as well as topic-oriented deep learning, which may also prove to be a valuable research direction.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61602036 and 61751201) and the National Key Research and Development Program of China (No.2017YFB0803302). The authors appreciate the anonymous reviewers for their valuable reviews.

References

[1] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1998, pp. 335–336.

[2] W. T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2007, pp. 1776–1782.

[3] K. Filippova, E. Alfonseca, C. A. Colmenares, L. Kaiser, and O. Vinyals, "Sentence compression by deletion with LSTMs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 360–368.

[4] L. Wang, H. Raghavan, V. Castelli, R. Florian, and C. Cardie, "A sentence compression based framework to query-focused multi-document summarization," in *Proc. ACL*, 2013, pp. 1384–1394.

[5] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer: A parse-and-trim approach to headline generation," in *Proc. HLT-NAACL 03 Text Summarization Workshop-Vol. 5*, 2003, pp. 1–8.

[6] Z. Cao, W. Li, S. Li, F. Wei, and Y. Li, "Attsum: Joint learning of focusing and summarization with neural attention," in *Proc. COLING*, 2016, pp. 547–556.

[7] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2016, pp. 484–494.

[8] W. Yin and Y. Pei, "Optimizing sentence modeling and selection for document summarization," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 1383–1389.

[9] P. Li, Z. Wang, W. Lam, Z. Ren, and L. Bing, "Salience estimation via variational auto-encoders for multi-document summarization," in *Proc. AAAI*, 2017, pp. 3497–3503.

[10] J.-g. Yao, X. Wan, and J. Xiao, "Recent advances in document summarization," *Knowl. Inf. Syst.*, vol. 53, no. 2, pp. 297–336, Nov. 2017.

[11] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: A survey," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 1–66, Jan. 2017.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. - Vol. 2*, 2013, pp. 3111–3119.

[14] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn. - Vol. 32*, 2014, pp. 1188–1196.

[15] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 3294–3302.

[16] M. Kågebäck, O. Mogren, N. Tahmasebi, and D. Dubhashi, "Extractive summarization using continuous vector space models," in *Proc. 2nd Workshop Continuous Vector Space Models Compositionality*, 2014, pp. 31–39.

[17] H. Kobayashi, M. Noguchi, and T. Yatsuka, "Summarization based on embedding distributions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1984–1989.

[18] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2418–2424.

[19] J. Law, H. H. Zhuo, J. He, and E. Rong, "LTSG: Latent topical skip-gram for mutually learning topic model and vector representations," *Chinese Conf. PRCV*, pp. 375–387, 2018.

[20] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai, "Jointly learning word embeddings and latent topics," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 375–384.

[21] R. Das, M. Zaheer, and C. Dyer, "Gaussian LDA for topic models with word embeddings," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics/7th Int. Joint Conf. Natural Lang. Process. (Vol. 1: Long Papers)*, 2015, pp. 795–804.

[22] S. Li, T.-S. Chua, J. Zhu, and C. Miao, "Generative topic embedding: A continuous representation of documents," in *Proc. ACL (1)*, 2016, pp. 666–675.

[23] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2177–2185.
[24] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 211–225, 2015.
[25] T. Kenter, A. Borisov, and M. D. Rijke, "Siamese CBOW: Optimizing word embeddings for sentence representations," in *Proc. ACL*, 2016, pp. 941–951.
[26] J. Nam, E. L. Mencía, and J. Fürnkranz, "All-in text: learning document, label, and word representations jointly," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1948–1954.
[27] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding" in *Proc. Interspeech*, 2013, pp. 3771–3775.
[28] A. Nenkova, L. Vanderwende, and K. Mckeown, "A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 573–580.
[29] X. Wan and J. Xiao, "Graph-based multi-modality learning for topic-focused multi-document summarization," in *Proc. 21st Int. Jont Conf. Artif. Intell.*, 2009, pp. 1586–1591.
[30] C. Y. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proc. 18th Conf. Comput. Linguistics - Vol. 1*, 2000, pp. 495–501.
[31] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2005, pp. 202–209.
[32] Y. Ouyang, W. Li, S. Li, and Q. Lu, "Applying regression models to query-focused multi-document summarization," *Inf. Process. Manage.*, vol. 47, no. 2, pp. 227–237, 2011.
[33] Y. Wang, H. Huang, C. Feng, Q. Zhou, J. Gu, and X. Gao, "CSE: Conceptual sentence embeddings based on attention model," in *Proc. ACL*, 2016, pp. 505–515.
[34] K.-Y. Chen, S.-H. Liu, H.-M. Wang, B. Chen, and H.-H. Chen, "Leveraging word embeddings for spoken document summarization," *INTERSPEECH*, pp. 1383–1387, 2015.
[35] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, "Learning summary prior representation for extractive summarization," in *Proc. ACL (2)*, 2015, pp. 829–833.
[36] D. Parveen, H. Ramsl, and M. Strube, "Topical coherence for graph-based extractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1949–1954.
[37] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," arXiv preprint cs/0405039, 2004.
[38] S. Gupta, A. Nenkova, and D. Jurafsky, "Measuring importance and query relevance in topic-focused multi-document summarization," in *Proc. 45th Annu. Meeting ACL Interactive Poster Demonstration Sessions*, 2007, pp. 193–196.
[39] J. Tang, L. Yao, and D. Chen, "Multi-topic based query-oriented summarization," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 1147–1158.
[40] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proc. SIGIR Res. Develop. Inf. Retrieval*, 2008, pp. 307–314.
[41] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2210–2216.
[42] M. Yang, T. Cui, and W. Tu, "Ordering-sensitive and semantic-aware topic modeling," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2353–2359.
[43] Z. Cao, W. Li, S. Li, and F. Wei, "Improving multi-document summarization via text classification," *AAAI*, pp. 3053–3059, 2017.
[44] J. P. Ng, P. Bysani, Z. Lin, M. Y. Kan, and C. L. Tan, "Swing: Exploiting category-specific information for guided summarization," in *Proc. TAC*, 2011.
[45] D. Gillick and B. Favre, "A scalable global model for summarization," in *Proc. Workshop Integer Linear Program. Natural Lang. Process.*, 2009, pp. 10–18.
[46] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proc. NAACL*, 2003, pp. 71–78.
[47] X. Wan and T. Wang, "Automatic labeling of topic models using text summaries," in *Proc. ACL*, vol. 1, 2016, pp. 2297–2305.

Yang Gao received the PhD degree from the Queensland University of Technology. Currently, she is a lecturer with the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China. Her current research interests include NLP, data mining, and machine learning. Her papers have been published in journals and conferences, such as the *IEEE Transactions on Knowledge and Data Engineering*, the *ACM Transactions on Intelligent Systems and Technology*, the *Knowledge and Information Systems*, WWW, AAAI, COLING, WISE, ICDM, and PAKDD.

Yue Xu is an associate professor from the Queensland University of Technology, Australia. She has worked in knowledge based systems for many years. Her current research interests include web intelligence and data mining. She has published more than 140 refereed papers covering research areas of association rule and pattern mining, recommender systems, and trust and reputation management, etc. Some of the papers have been published in major journals and conferences, such as *TKDE*, ICDM, CIKM, PAKDD, HT, WISE, WWW, and Web Intelligence.

Heyan Huang received the ME degree in computer science and technology from the National University of Defense Technology and the PhD degree from China Academy of Sciences, Institute of Computer Technology. Now, she is a professor and dean of the School of Computer Science and Technology, Beijing Institute of Technology, and the director of the Research Center of High Volume Language Information Processing and Cloud Computing. Her current research interests mainly focus on machine translation.

Qian Liu is working toward the PhD degree in the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include NLP and transfer learning. Her papers have been published in AAAI, COLING, and DASFAA.

Linjing Wei received the master's degree from the School of Computer Science and Technology, Beijing Institute of Technology, in 2017. Her research interests include text summarization and machine learning.

Luyang Liu is working toward the PhD degree in the School of Computer Science and Technology, Beijing Institute of Technology. His research interests include NLP and machine learning.