Information Sciences 607 (2022) 186–210

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A context-enhanced sentence representation learning method for close domains with topic modeling

Shuangyin Li [a,*], Weiwei Chen [b], Yu Zhang [c], Gansen Zhao [a], Rong Pan [b], Zhenhua Huang [a], Yong Tang [a]

[a] School of Computer Science, South China Normal University, Guangzhou, Guangdong, China
[b] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong, China
[c] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, Guangdong, China

ARTICLE INFO

Article history:
Received 31 January 2020
Received in revised form 28 May 2022
Accepted 29 May 2022
Available online 2 June 2022

Keywords:
Sentence representations learning
Closed domains
Bayesian sentence embedding
Bi-directional context-enhanced
Semantic interpretability
Topic modeling

ABSTRACT

Sentence representation approaches have been widely used and proven to be effective in many text modeling tasks and downstream applications. Many recent proposals are available on learning sentence representations based on deep neural frameworks. However, these methods are pre-trained in open domains and depend on the availability of large-scale data for model fitting. As a result, they may fail in some special scenarios, where data are sparse and embedding interpretations are required, such as legal, medical, or technical fields. In this paper, we present an unsupervised learning method to exploit representations of sentences for some closed domains via topic modeling. We reformulate the inference process of the sentences with the corresponding contextual sentences and the associated words, and propose an effective context-enhanced process called the bi-Directional Context-enhanced Sentence Representation Learning (bi-DCSR). This method takes advantage of the semantic distributions of the nearby contextual sentences and the associated words to form a context-enhanced sentence representation. To support the bi-DCSR, we develop a novel Bayesian topic model to embed sentences and words into the same latent interpretable topic space called the Hybrid Priors Topic Model (HPTM). Based on the defined topic space by the HPTM, the bi-DCSR method learns the embedding of a sentence by the two-directional contextual sentences and the words in it, which allows us to efficiently learn high-quality sentence representations in such closed domains. In addition to an open-domain dataset from Wikipedia, our method is validated using three closed-domain datasets from legal cases, electronic medical records, and technical reports. Our experiments indicate that the HPTM significantly outperforms on language modeling and topic coherence, compared with the existing topic models. Meanwhile, the bi-DCSR method does not only outperform the state-of-the-art unsupervised learning methods on closed domain sentence classification tasks, but also yields competitive performance compared to these established approaches on the open domain. Additionally, the visualizations of the semantics of sentences and words demonstrate the interpretable capacity of our model.

* Corresponding author.

https://doi.org/10.1016/j.ins.2022.05.113
0020-0255/

1. Introduction

Sentence representation learning is an important step in natural language process (NLP) tasks, and methods for learning meaningful representations of sentences have received significant attention in recent years. Retrieving a high-quality sentence representation has been a longstanding open problem at the intersection of language modeling and understanding. To date, numerous learning methods have been introduced to map sentences into embeddings, which can be divided into two categories: semi-supervised/supervised learning methods and unsupervised learning methods.

Semi-supervised/supervised learning methods depend on class labels and are tuned only for their respective tasks. This type of learning framework needs the labeled sentences to produce task-dependent embeddings [1], and the corresponding approaches are based on recursive networks [2], recurrent neural networks [3], and so on. For the unsupervised learning approaches, many researchers have built a bridge between the word embedding tasks and the sentence embedding tasks that learns sentence representations by encoding word vectors or imitates the inspiration of word embedding to the sentence level, for example of Skip-Thoughts [4] and Quick-Thoughts [5]. Recently, Bert [6] and the transformer-based methods, such as XLNET [7] and RoBERTa [8], have achieved a great success on many NLP tasks, which also provide an approach for sentence representation learning.

Despite the fast developmental pace of sentence embedding techniques, there are still two challenges that remain open. First, in many special scenarios, collecting enough data is laborious and time-consuming. It is difficult to obtain enough data to retrain a deep learning model or a transformer-based model from scratch. Though the pre-trained models provided by Bert or Quick-Thoughts on large-scale open-domain corpus are available, these fine-grained and pre-trained models usually have limitations to many closed-domain tasks in special scenarios, such as sentence classifications on legal, medical, or technical fields. For example, the well-known pre-trained model of Bert is trained on BooksCorpus and Wikipedia, and Quick-Thoughts is trained on BooksCorpus. Many words, such as proper nouns and entities in the special closed domain, are unseen by these pre-trained models which limits the transfer of the pre-trained model to the closed domain tasks. Second, the sentence representation interpretability becomes more crucial for certain real applications, especially the tasks from the medical domain. The distributed sentence embeddings generated by deep learning methods and transformer-based methods are weak on the semantic interpretation [9]. In many scenarios, we need to determine the proportions from a sentence representation on the semantic level for the similarity measures among sentences and words. Thus, it still is an open issue to train high-quality and interpretable sentence representations, especially in closed domains.

To capture the semantics of text, topic models are effective methods for text inference, and recently, many researchers have focused on text embedding with topic models [10]. The benefit of topic models is that the representations are highly interpretable with the Bayesian assumption on text and words [11–13]. As a probabilistic language modeling, the target of topic models is to group semantic-similar words to form a topic space, and represent text over the topic space. Meanwhile, topic models are available and effective in closed domains for document embedding where there isn't much call for large-scale data [14]. However, for the sentence representation learning task, existing topic models cannot address the problem of the lack of word information in sentences, because sentences are always short, which limits their application on sentence embedding [15]. In general, there are two main characteristics of sentences that need to be considered when a sentence embedding method is designed for closed domains. The first is that the sentences and words are highly semantic-similar, where the inner words of one sentence contain the main semantics of the sentence, especially the keywords. We can approximate the semantics of the sentence based on the meanings of these keywords. Second, the semantics of a sentence are highly related to its context, for example, the surrounding sentences. This property is used by many neural language models such as Quick-Thoughts and Bert. Through the surrounding sentences, we can obtain more information about the semantic environment and locate the topics the target sentence is involved in. It is clear that the contextual sentences in two directions can be used to enhance the semantics of a sentence, and they are critical to learning the sentence representation. Representation learning on sentence or short texts with topic modeling has received substantial attention in the literature [16–18,13]. However, the above two characteristics are not adopted by the existing techniques for the sentence or short text embedding on topic modeling.

Based on the two identified characteristics, we investigate the issue of forming a more meaningful representation of the target sentence solely by the semantics of the inner words and the contextual sentences in closed domains. In this paper, we propose a bi-Directional Context-enhanced Sentence Representation Learning (bi-DCSR) method to exploit representations of sentences via topic modeling by fully utilizing the sentence itself and the contextual sentences. That is, the bi-DCSR method takes a target sentence and its contextual sentences as input; the outputs are the context-enhanced representations by a context-enhanced process that is based on the semantic distributions of nearby contextual sentences and the associated words. In this process, we need to define sentences and words in the same topic space, which have not received consideration in most previous works.

To support the proposed bi-DCSR method, we develop a Hybrid Priors Topic Model (HPTM), which constructs the semantics of the sentences and words by learned topic distributions. The HPTM represents the sequences of sentences and it outputs three metrics: the topic distributions of the words in the dictionary, the topic distribution of each sentence, and the word probabilities of the latent topics. The HPTM aims to embed the sentences and the words into the same interpretable topic space, which is used in the context-enhanced process of the bi-DCSR method. The overall process of the bi-DCSR is shown as in Fig. 1.

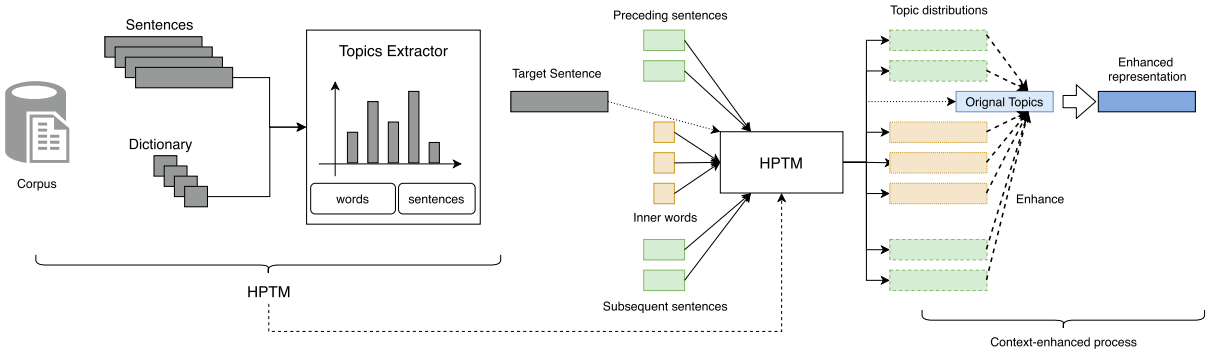


Fig. 1. The process of bi-DCSR to enhance the representation of a target sentence by the contextual sentences and the inner words in it.

We validate our model with three closed-domain datasets from legal cases, electronic medical records, and technical reports. In addition, we test the proposed method on one open-domain dataset from Wikipedia. We test the HPTM on language modeling and topic coherence with the four datasets, and the experiments show that the HPTM significantly outperforms existing topic models on language modeling and topic coherence. With respect to sentence classification tasks, the bi-DCSR method yields a state-of-the-art performance, and the results are significantly better than those obtained by state-of-the-art methods on closed domains. Meanwhile, we also test our model on Wikipedia to show the capacity of sentence embedding learning on an open domain, and the proposed model still achieves competitive results, especially compared with existing methods based on deep neural networks. Moreover, the bi-DCSR method is flexible as it can merge with other topic models on sentence representation learning, and the experiments suggest that it improves upon the performance of these topic models. The visualizations of sentence embeddings show the capacity of the proposed method to capture the semantic relatedness of both sentences and words, which is crucial for certain real applications on some closed domains, e.g., intention detection on medical dialog and Question–Answer systems.

The main contributions of this work can be summarized as follows.

1. To tackle the issue of sentence representation learning in closed domains, we propose a context-enhanced sentence representation learning method (bi-DCSR). The proposed bi-DCSR has much more effective training and inference processes. The proposed method is readily trained from scratch with limited amount of training data, which is suitable for closed domains.
2. Through a context-enhanced process, the proposed method fully takes advantage of the bi-Directional contextual information to learn high-quality sentence representations. Experiments have demonstrated that with the bi-Directional contextual information, the proposed bi-DCSR can achieve the state-of-the-art performances in sentence classification tasks on three closed-domain corpora.
3. To support the bi-DCSR method, we present a novel HPTM, a unified probabilistic language model of the sequences of sentences. The HPTM learns the topic distributions of sentences, topic distributions of all the words in the dictionary, and word probabilities of hidden topics. The proposed HPTM provides a way to embed words and sentences into the same topic space that is highly interpretable. To the best of our knowledge, this is the first work to embed sentences and words into the same interpretable space with topic modeling.
4. An online algorithm is also proposed to allow HPTM method be applied in open domains, by which the proposed bi-DCSR method is even competitive with systems tuned on open domain scenarios while also being extremely efficient and easy to use.

The rest of this paper is organized as follows. Section 2 introduces related works on sentence representation methods. Section 3 presents the proposed framework in detail, including bi-DCSR and HPTM. Section 7 reports experiments on sentence classification tasks and language modeling. Finally, Section 8 concludes this paper.

2. Related Works

Currently, there are many alternative proposals available for sentence representation learning. For example, it is a simple approach in many fields to average a sentence's word embeddings to be the sentence embedding is a simple approach in many fields, where the word embeddings are learned by different models [19–21]. Peters et al. introduce ELMo [19], a type of deep contextualized word representation through a bidirectional language model to capture context-dependent aspects of word meaning.

Another line of research uses a labeled or structured dataset to learn sentence representations with different NLP tasks [22]. These methods consider supervised tasks, such as machine translation, relation prediction, and training classifiers with man-made datasets to tune the models. Daniel et al. [22] present models for encoding sentences into embedding vectors that

specifically target at the capability of transfer learning to other NLP tasks. Wieting et al. [23] take advantage of paraphrase data to learn an encoder by learning general-purpose, paraphrastic sentence embeddings based on supervision from a paraphrase database.

Unsupervised learning approaches allow us to learn useful sentence representations from large unlabeled corpora [24]. Conneau et al. [9] introduce several probing tasks designed to capture simple linguistic features of sentences, and use them to uncover intriguing properties of both encoders and training methods. Pagliardini et al. [25] propose an unsupervised model for learning universal sentence embeddings with the sentence words being specifically optimized towards additive combinations over the sentence using compositional n-Gram features.

The idea of self-supervision is to learn sentence embeddings by designing suitable learning objectives of language models. These approaches are mainly based on encoder-decoder models [4] or autoencoder models [26]. Daniel et al. [27] present a sentence embedding model using a transformer and a deep averaging network as the sentence encoders. Zhao et al. [28] show an unsupervised discrete sentence representation learning method that can integrate with existing encoder-decoder dialog models for interpretable response generation. Based on the encoder-decoder architecture, these models take advantage of the word-level or sentence-level sequence information. They take the sentences as input and predict the neighbored sentences or words for decoding. Thus, it is necessary that the context information needs to be fully utilized, especially in unsupervised learning schemes.

Recently, transformer-based methods, such as Bert [6], XLNET [7], and RoBERTa [8], are used for sentence representation learning. For example, Bert-as-service[1] uses Bert as a sentence encoder to map sentences to fixed-length vector representations. Researchers pass single sentences through Bert and then derive a fixed-sized vector by either averaging the outputs (similar to average word embeddings) or by using the output of the special [CLS] token. However, this common practice yields poor sentence embeddings [29]. In most cases, it is worse than the averaging GloVe embeddings [20]. For example, May et al. [30] extend the word embedding association test to measure bias in sentence encoders, such as ELMo and Bert. Reimers et al. [29] present Sentence-BERT, which is a sentence embedding method using Siamese BERT-Networks. Other limitations of the transformer-based methods for sentence representation learning include the high cost of computing resources and the requirement of large-scale training data, which are not suitable to many closed-domains scenarios with only limited data available.

The work on topic models can also be used to learn sentence representations [31]. Topic models have been proven useful in document modeling, and many proposals have been made to handle corpora on word modeling [16], short text modeling [18,17], paragraph and document embedding [32], and other fields [33–35]. To date, many researchers still focus on the topic model and its variants [36], where the topic model is in a unique position to understand the interrelationships between words, sentences, and documents in the semantic space [10]. The topic models are often developed with a specific focus on deep neural networks [37]. However, many methods suffer from the lack of word information while learning sentence representations based on the words in the host sentences, due to the reason that sentences are always too short when they are compared with documents or paragraphs [15]. Li et al. [13] present a recurrent attentional topic model for the sentence level. Extracting topics from short texts or sentences is challenging because of severe data sparsity. Even though there is a demonstrated capacity of model interpretations using topic models, learning sentence representations based on topic models is still not well solved at present. The rich contextual information and the semantics of the inner words are helpful when building a sentence representation learning framework. Many works based on topic modeling ignore the relations of the sentences and the words in the semantic space.

Thus, to alleviate this problem, we propose a novel topic model that embeds sentences and words into the same topic space, which offers an opportunity to link the sentences and the inner words with latent topics. This approach can be used to build an effective sentence embedding method based on Bayesian topic modeling, overcoming the above limitations of topic modeling on sentence modeling. The proposed topic model focuses on the sentence level and a novel objective for sentence representation learning is designed to fully take advantage of the contextual information and the target sentence itself.

3. Bi-directional context-enhanced sentence representation learning

We aim to design a Bayesian probabilistic framework to learn the representations of the sentences enhanced by the bi-directional contexts and the inner words.

3.1. Notation

Let $D = \{\mathbf{d}^1, \mathbf{d}^2, \cdots, \mathbf{d}^M\}$ denote a corpus which contains M documents, where $\mathbf{d}^i, i \in \{1, \cdots, M\}$ indicates the i-th document in this corpus. Each document $\mathbf{d}^i$ in the corpus is defined as a sequence of sentences which denoted by $(\mathbf{s}^i_1, \mathbf{s}^i_2, \cdots, \mathbf{s}^i_{S^i})$, where S^i is the number of the sentences in $\mathbf{d}^i$. Similarly, each $\mathbf{s}^i_j$ is denoted by $(w^i_{j1}, w^i_{j2}, \cdots, w^i_{jN^i_j})$, where N^i_j is the number of words in $\mathbf{s}^i_j$. Here we adopt the bag-of-words assumption. A dictionary v is made up of the words in the corpus D, where the word is indexed by $\{1, 2, \cdots, V\}$. The notations are summarized in Table 1.

[1] https://github.com/hanxiao/bert-as-service

Table 1
Notations.

Symbol	Description
D	set of corpus which contains documents.
$\mathbf{d}^i$	the i-th document in the corpus D.
S^i	the total number of the sentences in $\mathbf{d}^i$, the i denotes the index.
$\mathbf{s}_j^i$	the j-th sentence in $\mathbf{d}^i$, where $\mathbf{d}^i$ is treated as a sequence of sentences.
N_j^i	the number of words in $\mathbf{s}_j^i$, where i denotes the index of the host document, and j denotes the index of the host sentence.
$\{w_{jn}^i\}$	the words in sentence $\mathbf{s}_j^i$, where n denotes the index of word w in sentence $\mathbf{s}_j^i$.
v	the dictionary of corpus D, which contains all the word appeared in D.
V	the size of v.
z_k	the topic index of topic k. The total number of topics in a sentence is K, and z_k denotes the k-th topics.
$\vartheta^{s_j^i}$	the topic distribution of sentence $\mathbf{s}_j^i$ by bi-DCSR. It is the context-enhanced representation of sentence s_j^i.
α	the hyperparameter used in LDA.
ζ	the hyperparameter used in HPTM.
$Dir(\cdot)$	Dirichlet distribution.
$F(\cdot)$	the weighted function that mixtures the topics, which defines the topic generation process.
ϕ_{w_n}	the topic distribution of word w_n. Each word defines a distribution over topics as well as the sentences.
C	the contextual window size used in HPTM.
C_r	the contextual window size used in bi-DCSR.
$\theta^{s_{j-}}$	the topic distribution matrices of the preceding sentences for $\mathbf{s}_j^i$.
$\theta^{s_{j-C_{j-1}}}$	the $C \times 1$ topic distribution matrices of the preceding sentences for $\mathbf{s}_j^i$.
$\theta^{s_{j+}}$	the topic distribution matrices of the subsequent sentences for $\mathbf{s}_j^i$.
$\theta^{s_{j+1j+C}}$	the $C \times 1$ topic distribution matrices of the subsequent sentences for $\mathbf{s}_j^i$.
ϵ	the weight vector for the contextual sentences and the words in generating the topic distribution of sentence $\mathbf{s}_j^i$, normally, the dimension of ϵ is $N + 2C$.
π	the parameter for ϵ, which follows a Dirichlet.
β	the word probabilities of the topics over the dictionary.
μ	the parameter for β, which follows multinomial distributions. The dimension of μ is V.
$Mult(\cdot)$	multinomial distribution
$\{\gamma_n\}_j^i$	a group of variational parameters of multinomial distributions for $\{z_{w_n}\}_j^i$, where $\{z_{w_n}\}_j^i$ denotes the topic index of w_n in sentence $\mathbf{s}_j^i$.
ξ_j^i	the variational parameter for ϵ_j^i, which follows a Dirichlet.
$\Psi(\cdot)$	the digamma function.

3.2. Topic Generation of Sentences with the Inner Words

With traditional LDA methods, the topic distribution of a sentence is generated from a Dirichlet distribution with a hyperparameter α as shown in Fig. 2(a). Let $T = (z_1, z_2, \cdots, z_K)$ to be the topic space, which can be referred to the LDA method. We define $\vartheta^{s_j^i}$ as the topic probability distribution of the j-th sentence $\mathbf{s}_j^i$ in $\mathbf{d}^i$. The sentence representations are viewed as the topic proportions over the topic space T.

As discussed above, keywords mainly determine the topic proportions in a sentence. For the convenience of description, we treat all of the inner words of the sentence as the keywords. The topic distribution of the sentence can be generated by the mixture of the topic distributions of the words, which is different from the assumption of the LDA method. We assume that the generation of a sentence begins with a mixture of topic distributions of the inner words; subsequently, the topic index of each word, z, is sampled from $\vartheta^{s_j^i}$. Fig. 2(b) indicates the topic mixture processes for $\vartheta^{s_j^i}$. Based on this assumption, we begin with the definition of the conditional probability of the sentence's topic distribution given the associated words.

Definition 1. Consider $\mathbf{s}_j^i$ and hidden variable $z \sim Dir(\cdot), z_k \in (z_1, z_2, \cdots, z_K)$, the posterior probability of the sentence's topic distribution is $p(\vartheta^{s_j^i}|w_{j1}^i, \cdots, w_{jN_j^i}^i; z_k) \propto p(\vartheta^{s_j^i}|z_k)p(z|w_{j1}^i, \cdots, w_{jN_j^i}^i)$, where $p(\cdot|z_k)$ is the class-conditional marginal distribution over a $K-1$ dimensional simplex, which we call factors[2] [38].

Let $p(\vartheta^{s_j^i}|w_{j1}^i, \cdots, w_{jN_j^i}^i; z_k)$ denote the posterior probability of the topic distribution of $\mathbf{s}_j^i$. Given the observed variables $w_{j1}^i, \cdots, w_{jN_j^i}^i$, we compute the marginal distribution of $\vartheta^{s_j^i}$ with topic z_k using Bayesian theorem:

$$p(\vartheta^{s_j^i}|w_{j1}^i, \cdots, w_{jN_j^i}^i; z_k) = p(\vartheta^{s_j^i}|z_k) \cdot p(z_k|w_{j1}^i, \cdots, w_{jN_j^i}^i) \cdot p(w_{j1}^i, \cdots, w_{jN_j^i}^i) \propto p(\vartheta^{s_j^i}|z_k) \cdot p(z_k|w_{j1}^i, \cdots, w_{jN_j^i}^i), \tag{1}$$

[2] Here we use the notation of $p(\cdot|z_k)$ to denote the topic proportion over z_k, the same as PLSA [38]

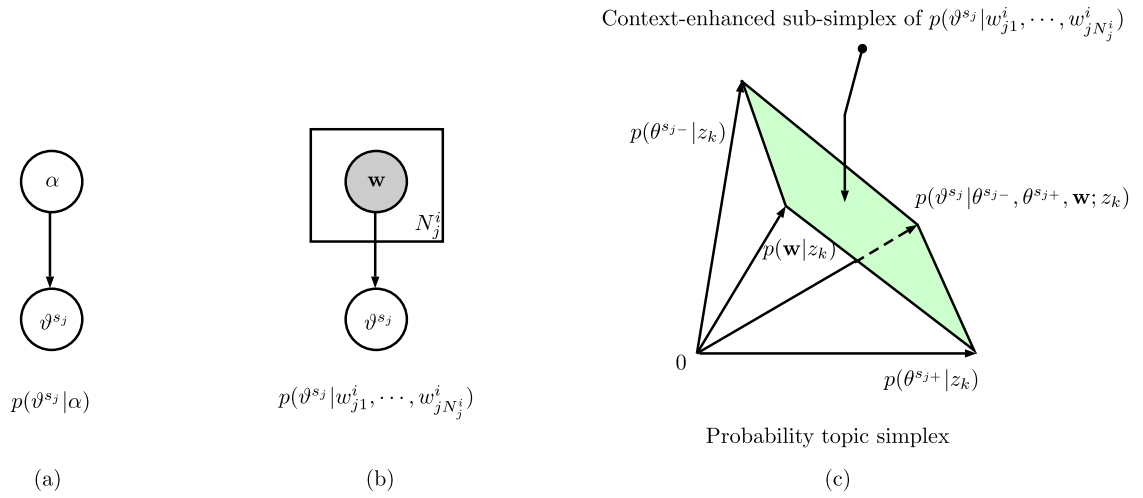


Fig. 2. (a): Topic generation for LDA on the sentence level. (b): Topic generation process of a sentence with the inner words as the priors. (c): The sketch of the context-enhanced sub-simplex for sentence $\mathbf{s}_j^i$ spanned by bi-DCSR. We let $\vartheta^{s_j^i}$ denote context-enhanced topic distribution. θ in $p(\theta|\cdot)$ denotes the topic distribution of other textual elements, such as a neighbored sentence. The bi-DCSR places a smooth distribution for sentence ϑ^{s_j} on the topic simplex denoted by the green surface.

where $p(w_{j1}^i,\cdots,w_{jN_j^i}^i)$ is a constant with respect to the observed $\mathbf{s}_j^i$. $p(z_k|w_{j1}^i,\cdots,w_{jN_j^i}^i)$ is the marginal distribution of z_k over the observed words $w_{j1}^i,\cdots,w_{jN_j^i}^i$ as

$$p(z_k|w_{j1}^i,\cdots,w_{jN_j^i}^i)=\sum_{n=1}^{N_j^i}p(z_k|w_{jn}^i).$$

Thus, the posterior probability of the sentence $\mathbf{s}_j^i$ on topic z_k can be computed as

$$p(\vartheta^{s_j^i}|w_{j1}^i,\cdots,w_{jN_j^i}^i;z_k)=\frac{p(\vartheta^{s_j^i}|z_k)\cdot\sum_{n=1}^{N_j^i}p(z_k|w_{jn}^i)}{\sum_{z\prime\in T}p(\vartheta^{s_j^i}|z\prime)\cdot\sum_{n=1}^{N_j^i}p(z\prime|w_{jn}^i)}. \tag{2}$$

Fig. 2(b) shows the graphical model of the posterior probability of $\vartheta^{s_j^i}$ given the inner words $w_{j1}^i,\cdots,w_{jN_j^i}^i$.

3.3. Enhance Topics of Sentences with the Contexts

To enhance the impact of the contexts, we delve deeper and reformulate the term $p(\vartheta^{s_j^i}|z_k)$ in Eq. (1) and Eq. (2) with a Context-Enhanced Process. There are two types of contexts that are brought into the term $p(\vartheta^{s_j^i}|z_k)$. (1) The bi-Directional contextual sentences of sentence $\mathbf{s}_j^i$. (2) The inner words of $\mathbf{s}_j^i$, which are treated as one type of contexts for the target sentence.

Definition 2. Given the topic distributions of the contextual sentences and the inner words, the enhanced topic distribution of $\mathbf{s}_j^i$ can be reformulated as follows:

$$p\left(\vartheta^{s_j^i}|z_k\right)_{\{\theta^{s_{j-}},\theta^{s_{j+}},w_{j1}^i,\cdots,w_{jN_j^i}^i\}}=p(\theta^{s_{j-}}|z_k)\cdot p(\theta^{s_{j+}}|z_k)\cdot p\left(\vartheta^{s_j^i}|\theta^{s_{j-}},\theta^{s_{j+}},w_{j1}^i,\cdots,w_{jN_j^i}^i;z_k\right)\cdot\prod_{n=1}^{N_j^i}p\left(w_{jn}^i|z_k\right),$$

where we introduce the following distributions:

(1) $\theta^{s_{j-}}$ denotes the topic distributions of the preceding sentences of $\mathbf{s}_j^i$ over T.
(2) $\theta^{s_{j+}}$ denotes the topic distributions of the subsequent sentences of $\mathbf{s}_j^i$ over T.

(3) $p(w^i_{jn}|z_k)$ denotes the topic distribution of word w^i_{jn} over T.

(4) $p(\vartheta^{s^i_j}|\theta^{s_{j-}}, \theta^{s_{j+}}, w^i_{j1}, \cdots, w^i_{jN^i_j})$ denotes the marginal distribution of $\vartheta^{s^i_j}$

with a group of priors $\{\theta^{s_{j-}}, \theta^{s_{j+}}, w^i_{j1}, \cdots, w^i_{jN^i_j}\}$ over T.

Definition 2 shows the context-enhanced process of modeling a sentence, which is an affine function that we can optimize to obtain the context-enhanced topic distribution of a sentence. Note that, $\theta^{s_{j-}}$ and $\theta^{s_{j+}}$ denotes the topic distributions of the neighbored sentences, and $\vartheta^{s^i_j}$ indicates the topic distribution of the current sentence for the purpose of distinguishing from the contextual sentences. All the topic distributions are defined in the same $K-1$-dimensional simplex. Likely, $p(w^i_{jn}|z_k)$ indicates the topic distribution of w^i_{jn}, which is defined over the topic space T as well as the sentences. In Definition 2, the preceding sentences, the subsequent sentences, and the inner words are all included in generating the topics of the current sentence. The graphical representation is shown in Fig. 2(c). We call this the bi-Directional Context-enhanced Representation (bi-DCSR) method, which is a Bayesian process that takes full advantage of the context of sentences.

The sentence representation learning process is illustrated by considering the geometry of Definition 2 and seeing how a sentence is represented by the bi-DCSR method. We introduce the topic simplex [38] and each topic distribution can be viewed as a point on the $(K-1)$-dimensional simplex. This process consists of three main steps and is highlighted in Fig. 2(c).

First, all of the words in the dictionary are drawn from a distribution over topics, i.e., points on the topic simplex, which means that the words are defined by probability distributions on T. Second, all of the sentences in the corpus are also drawn from a sentence-specific topic distribution, which are the points on the same topic simplex as the words. Third, the bi-DCSR method posits that the topic distribution of each sentence is drawn from the distribution defined in Definition 2, where the topic distributions of the preceding sentences, the subsequent sentences, and the inner words are also points from the same topic simplex. From Fig. 2(c), the modeling assumption expressed by Definition 2 is that conditional distributions $p(\vartheta^{s^i_j}|w^i_{j1}, \cdots, w^i_{jN^i_j})$ for all of the sentences are approximated by a multinomial distribution as a convex combination of factors including $p(\mathbf{w}|z_k), p(\theta^{s_{j+}}|z_k), p(\theta^{s_{j-}}|z_k)$ and $p(\theta^{s_j}|\theta^{s_{j-}}, \theta^{s_{j+}}, w^i_{j1}, \cdots, w^i_{jN^i_j}; z_k)$.

Given Definition 1 and Definition 2, the posterior probability of $\vartheta^{s^i_j}$ on topic z_k enhanced by the contexts and the inner words can be computed as

$$p(\vartheta^{s^i_j}|w^i_{j1}, \cdots, w^i_{jN^i_j}; z_k) \propto p(\vartheta^{s^i_j}|\theta^{s_{j-}}, \theta^{s_{j+}}, w^i_{j1}, \cdots, w^i_{jN^i_j}; z_k) \cdot p(\theta^{s_{j-}}|z_k) \cdot p(\theta^{s_{j+}}|z_k) \cdot \prod_{n=1}^{N^i_j} p(w^i_{jn}|z_k) \cdot \sum_{n=1}^{N^i_j} p(z_k|w^i_{jn}). \tag{3}$$

The proposed bi-DCSR method takes full advantage of the contexts in the topic modeling for sentence representation learning. The operation of embedding all of the text elements into the same topic simplex gives us an effective way to enhance the sentence representation by the contextual information, such as the keywords, preceding sentences, and subsequent sentences, in addition to paragraphs and phrases. This model, which embeds other text elements into the same topic simplex as the sentences, can extract high-quality sentence embeddings.

From the form of Eq. (3), we need to design a unified probabilistic language model of the sequences of sentences, that learns all of the elements appearing in the bi-DCSR method: the topic distributions of all of the sentences ($p(\theta^s|z_k)$), the topic distributions of all of the words in the dictionary ($p(\mathbf{w}|z_k)$), and the word probabilities of the hidden topics ($p(z_k|\mathbf{w})$). More specifically, the marginal distribution of $\vartheta^{s^i_j}, p(\vartheta^{s^i_j}|\theta^{s_{j-}}, \theta^{s_{j+}}, w^i_{j1}, \cdots, w^i_{jN^i_j})$ indicates a generation process of a sentence, which can be handled with topic modeling. Thus, we design a topic model learned with an unsupervised method to estimate the distributions utilized in the bi-DCSR method. This is presented in the next section.

4. Hybrid Priors Topic Model

To support the bi-DCSR method, we introduce a novel Bayesian topic model at the sentence level, HPTM, which jointly learns all three metrics: the topic distributions of the sentences, the topic distributions of the words in the dictionary, and the word probabilities of the hidden topics. Meanwhile, following the marginal distribution $p(\vartheta^{s^i_j}|\theta^{s_{j-}}, \theta^{s_{j+}}, w^i_{j1}, \cdots, w^i_{jN^i_j})$, the HPTM is tailored to model $\vartheta^{s^i_j}$ with bi-directional sequences of sentences and the inner words utilized in Definition 2. In general, the HPTM takes the bi-directional sequences of sentences as inputs, and aims to embed the sentences and the words into the same topic space, where the topic space is defined over the dictionary.

4.1. HPTM Definition

We let ϕ_{w_n} denote the topic distribution of the n-th word in sentence s. Based on the bi-directional context-aware assumption, we define a new Bayesian Process for generating the topic distribution of the sentence as

$$\vartheta^{s_j}|\theta^{s_{j-C:j-1}}, \theta^{s_{j+1:j+C}}, \phi_{w_1}, \cdots, \phi_{w_N}, \epsilon \sim \sum_{p}^{C} \epsilon_p \cdot \theta^{s_{j-p}} + \sum_{q}^{C} \epsilon_{q+C} \cdot \theta^{s_{j+q}} + \sum_{n}^{N} \epsilon_{n+2C} \cdot \phi_{w_n}, \tag{4}$$

where $\theta^{s_{j-C:j-1}}$ and $\theta^{s_{j+1:j+C}}$ denote the topic distribution matrices of the preceding and subsequent sentences, respectively. C is the contextual window size. $\phi_{w_1}, \cdots, \phi_{w_N}$ indicate the topic distribution matrices of the words appearing in sentence $\mathbf{s}_j^i$. ϵ is an $(N+2C) \times 1$-dimensional weight vector, which follows a Dirichlet distribution with a hyperparameter π.

The above Bayesian process defines a marginal distribution of $\vartheta^{s_j^i}$, $p(\vartheta^{s_j^i}|\theta^{s_{j-}}, \theta^{s_{j+}}, w_{j1}^i, \cdots, w_{jN_j^i}^i)$, over the preceding and subsequent sentences and the inner words with the corresponding weight values. We use

$F(\theta^{s_{j-C:j-1}}, \theta^{s_{j+1:j+C}}, \phi_{w_1}, \cdots, \phi_{w_N}, \epsilon)$ to denote the mixture function of the priors, and the output of $F(\cdot)$ is a K-dimensional vector. The prior of ϑ^s is a hybrid of the contextual sentences and the inner words with the Bayesian process of $F(\cdot)$ shown as in Fig. 3.

Proposition 1. Consider $\epsilon \sim Dir(\pi), \phi_{w_n} \sim Dir(\zeta), \epsilon \in (0,1)^{1\times(2C+N)}, \phi_{w_n} \in (0,1)^{1\times K}, \vartheta^{s_j^i}$ satisfies $\sum_k \vartheta_k^{s_j^i} = 1$, and $0 \leqslant \vartheta_k^{s_j^i} \leqslant 1$, where $\vartheta^{s_j^i}$ follows a Bayesian Process defined by $\vartheta^{s_j^i} \sim \sum_p^C \epsilon_p \cdot \theta^{s_{j-p}} + \sum_q^C \epsilon_{q+C} \cdot \theta^{s_{j+q}} + \sum_n^N \epsilon_{n+2C} \cdot \phi_{w_n}$.

Proof 1. Since $\epsilon \sim Dir(\pi), \epsilon$ satisfies $\sum_i \epsilon_i = 1$, and $0 \leqslant \epsilon_i \leqslant 1$ according to the property of Dirichlet distribution. Also, $\phi_{w_n} \sim Dir(\zeta)$ satisfies $\sum_k \phi_{w_n}^k = 1$, and $0 \leqslant \phi_{w_n}^k \leqslant 1$. The form of $\vartheta^{s_j^i}$ follows a recursive definition, where the topic distributions of the preceding and subsequent sentences satisfy $\sum_k \theta_k^{s_{j-}} = 1, 0 \leqslant \theta_k^{s_{j-}} \leqslant 1$ and $\sum_k \theta_k^{s_{j+}} = 1, 0 \leqslant \theta_k^{s_{j+}} \leqslant 1$. Thus, $\vartheta^{s_j^i}$ satisfies $\sum_k \vartheta_k^{s_j^i} = 1$, and $0 \leqslant \vartheta_k^{s_j^i} \leqslant 1$.

Based on Proposition 1, the random variable of $\vartheta^{s_j^i}$ with a hybrid prior generated by $F(\cdot)$ also follows a Dirichlet distribution, which enables the modeling of the sentences and words in the same topic simplex space [39]. Note that the dimension of ϵ is specified by the host sentence. The hyperparameter ϵ of the Dirichlet distribution is selected from a global vector $\pi \in \mathbb{R}^{1\times(2C+V)}$ corresponding to the word index, where V is the size of the dictionary.

Based on the definition of the topic distribution for a sentence, we can describe the generation process of the HPTM as follows:

1. For each hidden topic $k \in \{1, \ldots, K\}$, draw $\beta_k \sim Dir(\mu)$, where μ is a V-dimensional parameter vector;
2. For each word $w_v \in \{1, \ldots, V\}$in the dictionary, draw $\phi_{w_v} \sim Dir(\zeta)$, where ζ is a K-dimensional parameter vector;
3. For current sentence $\mathbf{s}_j, j \in \{1, \ldots, S_i\}$ in the document $\mathbf{d}^i$:
   (a) Draw $\vartheta^{s_j^i}$ following Eq. (4).
   (b) For each word $w_n, n \in \{1, \ldots, N_j\}$ in sentence $\mathbf{s}_j$:
      (i) Draw $z_n \sim Mult(\vartheta^{s_j})$;
      (ii) Draw $w_n \sim Mult(\beta_{z_n})$.

In this process, $Mult(\cdot)$ is a multinomial distribution. $\beta \in \mathbb{R}^{K\times V}$ denotes word probabilities of the hidden topics over the dictionary in the corpus, and each row in $\phi \in \mathbb{R}^{V\times K}$ is the topic distribution of one word in the dictionary. From this generation process, the topic distribution of each sentence is determined by its bi-directional contexts and the words in it. After learning this model, we can obtain β, ϕ, and the topic distribution of each sentence. Fig. 4 shows the graphical representation of the HPTM.

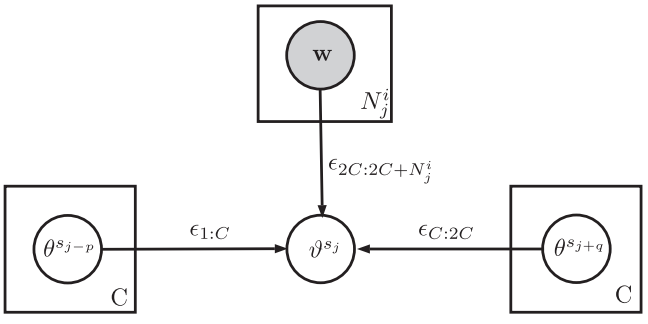


Fig. 3. The illustration of Bayesian process for generating the topic distribution of sentence $\mathbf{s}_j^i$ by the preceding and subsequent sentences and the associated words with the corresponding weight values.

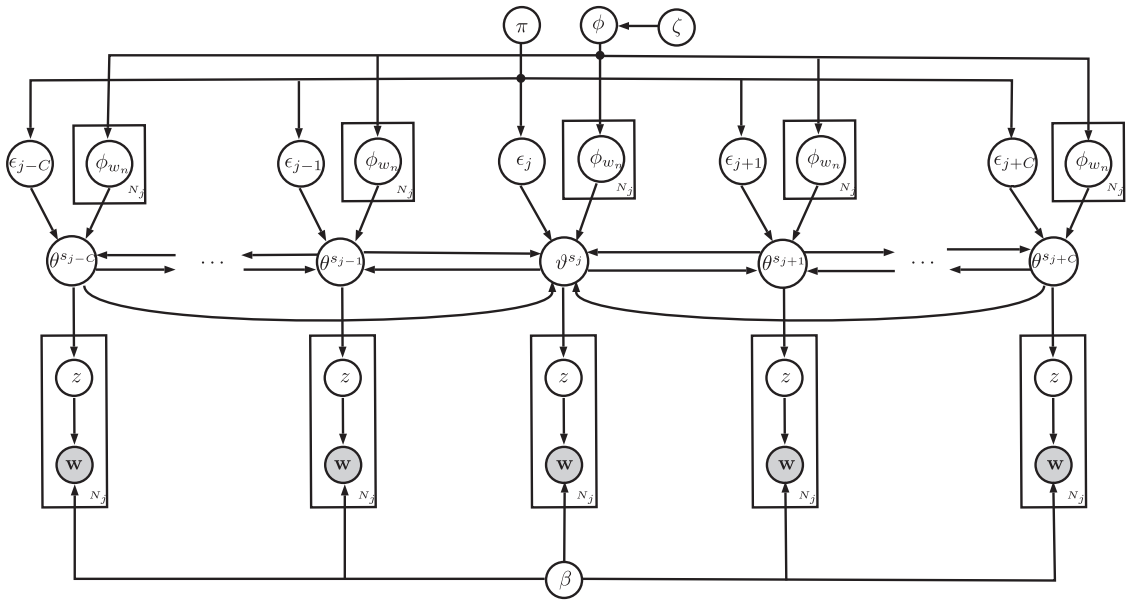


Fig. 4. Graphical representation of the HPTM.

Many works based on topic modeling have been proposed to handle sequences of sentences [13]. To the best of our knowledge, the proposed HPTM is a novel sentence-level topic model for embedding sentences and words into the unified semantic space. Considering the recurrent attentional topic model(RATM) [13] as an example, we see that it takes advantage of sequences of sentences and generates the target sentence by its preceding sentences. Different from the RATM, the HPTM leverages the two-directional contextual information and the inner words to learn the topics. To handle the data sparsity issue in short documents, some topic models either assume that a short document only covers a single topic or assume that the words in each sentence are drawn from the same topic. The HPTM relaxes these constraints by bringing richer contextual information and fully utilizing co-occurrences of the associated words. This approach supports the bi-DCSR method in that the context-enhanced process requires the contextual information to be represented in a unified semantic space.

4.2. Estimation of the HPTM

Since the topic generation process utilizing $F(\cdot)$ leads to the non-conjugate relation between the topic assignment and the prior over the topic distribution of sentences, we adopt to the variational Bayesian inference.

The latent variable of sentence $\mathbf{s}_j^i$ is the topic assignment z_{w_n} for each word and the weight vector ϵ_j^i. We define $\{\gamma_n\}_j^i$ as a group of variational parameters of multinomial distributions for $\{z_{w_n}\}_j^i$ and let $\xi_j^i \in \mathbb{R}^{(N+2C)\times 1}$ denote the variational parameter of a Dirichlet distribution for ϵ_j^i. Thus, for sentence $\mathbf{s}_j^i$, we use the following fully-factorized variational distribution

$$q(\epsilon_j^i, \{z_{w_n}\}_j^i) = q(\epsilon_j^i|\xi_j^i)\prod_{n=1}^{N_j^i} q(z_{w_n}|\gamma_n). \tag{5}$$

Thus, based on Jensen's inequality, the lower bound on the log probability of the sentence $\mathbf{s}_j^i$ given the model parameters $\{\beta, \phi, \pi\}$ can be computed as

$$\begin{aligned}\mathscr{L}^s(\epsilon, \gamma; \beta, \phi, \pi) = {} & \mathbb{E}_q[\log p(\epsilon|\pi)] + \sum_{n=1}^{N_j^i}\mathbb{E}_q[\log p(z_n|\epsilon, \theta^{s_{j-C:j-1}}, \theta^{s_{j+1:j+C}}, \phi_{w_1}, \cdots, \phi_{w_N})] + \sum_{n=1}^{N_j^i}\mathbb{E}_q[\log p(w_n|z_n, \beta)] \\ & - \mathbb{E}_q[\log p(\epsilon)] - \mathbb{E}_q[\log p(z_n)],\end{aligned} \tag{6}$$

where the last two terms indicate the entropy of the variational distributions. Based on the above fully factorized variational distribution, we can maximize the evidence lower-bound (ELBO) to find the solution of the variational parameters via the variational expectation–maximization (EM) procedure, where E-step and M-step are included. For the variational parameter ξ, the corresponding objective function is to minimize the following function

$$\mathscr{L}^{s}_{[\xi]} = \sum_{l}^{2C+N^i_j}(\pi_l - 1)(\Psi(\xi_l) - \Psi(\sum_{l\prime=1}^{2C+N^i_j}\xi_{l\prime})) - \log\Gamma(\sum_{l=1}^{2C+N^i_j}\xi_l) + \sum_{l=1}^{2C+N^i_j}\log\Gamma(\xi_l) - \sum_{l=1}^{2C+N^i_j}(\xi_l - 1)\cdot\left(\Psi(\xi_l) - \Psi\left(\sum_{l\prime=1}^{2C+N^i_j}\xi_{l\prime}\right)\right) + \sum_{n=1}^{N^i_j}\sum_{k=1}^{K}\gamma_{nk}\cdot\left(\sum_{l=1}^{2C}\log\theta^{j-C:j+C}_{lk}\frac{\xi_l}{\sum\xi} + \sum_{l=2C}^{2C+N^i_j}\phi_{w_n}\frac{\xi_l}{\sum\xi}\right), \tag{7}$$

where $\Psi(\cdot)$ indicates the digamma function that is the first derivative of the log of the Gamma function.

Also, after setting the derivatives of the ELBO to zero, we obtain the update equations for $\{\gamma_n\}^i_j$ as

$$\gamma_{nk} \propto \beta_{k,v^{w_n}}\exp\left(\sum_{l=1}^{2C}\log\theta^{j-C:j+C}_{lk}\frac{\xi_l}{\sum\xi} + \sum_{l=2C}^{2C+N^i_j}\phi_{w_n}\frac{\xi_l}{\sum\xi}\right), \tag{8}$$

where v^{w_n} denotes the index of the word w_n in the dictionary. Thus, we can optimize the variational parameters, i.e., ξ and $\{\gamma_n\}^i_j$, for each sentence in the E-step.

In M-step, we update the model parameters $\{\beta, \phi, \pi\}$ by maximizing the lower bound after fitting ξ and $\{\gamma_n\}^i_j$. The update equations for ϕ and β are as follows:

$$\phi_{vk} \propto \sum_{i}^{M}\sum_{j}^{S^i}\gamma_{v^{w_n}k}\frac{\xi_{v^{w_n}}}{\sum_{l=2C}^{2C+N^i_j}\xi_l}, \tag{9}$$

and

$$\beta_{kv} = \sum_{i=1}^{M}\sum_{j=1}^{S^i}\sum_{n=1}^{N^i_j}\gamma_{v^{w_n}k}(w_n)^v. \tag{10}$$

We use a gradient descent method to update π whose objective function is to minimize the following function

$$\mathscr{L}_{[\pi]} = \sum_{i=1}^{M}\sum_{j=1}^{S_i}(\log\Gamma(\sum_{l=1}^{2C+N^i_j}\pi_l) - \sum_{l\prime=1}^{2C+N^i_j}\log\Gamma(\pi_{l\prime}) + \sum_{l=1}^{2C+N^i_j}(\pi_l - 1)(\Psi(\xi_l) - \Psi(\sum_{l\prime=1}^{2C+N^i_j}\xi_{l\prime}))). \tag{11}$$

We use the linear-time Newton–Raphson algorithm to estimate π. The whole algorithm is shown in Algorithm 1.

Algorithm 1: Variational EM algorithm for HPTM.

1: INPUT: Sentences $(\mathbf{s}^i_1, \mathbf{s}^i_2, \cdots, \mathbf{s}^i_{S^i})$ in corpus C
2: OUTPUT: The topic distributions of the sentences and words, model parameters $\{\beta, \phi, \pi\}$.
3: **repeat**
4: *(E-Step:)*
5: **for**each sentence $\mathbf{s}^i_j$ of each document $\mathbf{d}^i$ in **Ddo**
6: update ξ^i_j via Eq. (7)
7: update γ_{nk} via Eq. (8).
8: **end for**
9: *(M-Step:)*
10: update ϕ_{vk} via Eq. (9)
11: update β via Eq. (10)
12: update π via Eq. (11).
13: **until**convergence

The E-Step in Algorithm 1 is to estimate the variational parameters for each sentence in the documents, and the M-Step is to learn the model parameters in the current iteration, which is the standard algorithmic process to maximize the ELBO to find the solution of the model parameters via the variational expectation–maximization procedure. Step 6 and Step 7 in Algorithm 1 update the variational parameters of ξ^i_j and γ_{nk} to find the sufficient statistics with respect to the model parameters. Step 10, 11 and 12 in M-Step is to estimate the model parameters $\{\phi, \beta, \pi\}$, respectively. The algorithm is convergent when $\{\phi, \beta, \pi\}$ satisfy the convergence criteria.

4.3. Online Variational Bayesian Inference for the HPTM

In order to handle the sentence streams in many domains, we also present an online variational Bayesian inference [40,41] for HPTM, which considers mini-batches of data in each update to reduce noise. To accomplish this, we split the corpus **D** into a sequence of mini-batches $\mathbf{b} \in (0, \infty)$. The online variational Bayesian inference is based on the fully-factorized variational distributions in Eq. (5); ϕ_{vk}, β and π are updated with each mini-batch in the M-step.

Given a batch **b** with B documents, we can obtain the following update equation for ϕ

$$\phi_{vk} \propto \sum_{i}^{B}\sum_{j}^{S^i}\gamma_{v^{w_n}k}\frac{\xi_{v^{w_n}}}{\sum_{l=2C}^{2C+N_j^i}\xi_l}. \tag{12}$$

For a new batch, we need update ϕ using a weighted average of its previous values ϕ_{old} and the new value ϕ_b in the current batch **b**. The weight given to ϕ is defined as

$$\psi^b = (\tau_0 + b)^{-\eta}, \tau_0 \geqslant 0,$$

where, $\eta \in (0.5, 1]$ controls the rate at which previous values are forgotten as used in [42].

Thus, after computing the gradient by $\nabla\phi = \phi_{old} - \phi_b$, we can update ϕ by a stochastic natural gradient algorithm as

$$\phi = (1 - \psi^b) \cdot \phi_{old} + \psi^b \cdot \phi_b. \tag{13}$$

Similarly, β can be also updated as

$$\beta_{kv} = (1 - \psi^b) \cdot \beta_{kv} + \psi^b \cdot \frac{D}{B}\sum_{i=1}^{B}\sum_{j=1}^{S^i}\sum_{n=1}^{N_j^i}\gamma_{v^{w_n}k}(w_n)^v, \tag{14}$$

where D is the total number of documents in the population. In the online setting, this can be an estimation of the maximum number of documents that may ever be seen.

For each batch, we use the gradient descent method to update π whose objective function is to minimize the following function

$$\mathscr{L}_{[\pi]} = \sum_{i=1}^{B}\sum_{j=1}^{S_i}(\log\Gamma(\sum_{l=1}^{2C+N_j^i}\pi_l) - \sum_{l'=1}^{2C+N_j^i}\log\Gamma(\pi_{l'}) + \sum_{l=1}^{2C+N_j^i}(\pi_l - 1)(\Psi(\xi_l) - \Psi(\sum_{l'=1}^{2C+N_j^i}\xi_{l'}))). \tag{15}$$

The whole online algorithm is shown in Algorithm 2.

Algorithm 2: Online variational EM algorithm for HPTM.

```
INPUT: Sentences (s_1^i, s_2^i, ..., s_{S^i}^i) in corpus C
OUTPUT: The topic distributions of the sentences and words, model parameters.
Define ψ^b = (τ_0 + b)^{-η}.
for b = 0 to ∞ do
  (E-Step:)
  repeat
    for each sentence s_j^i of each document d^i in b do
      update ξ_j^i, γ_nk via Eq. (7) and Eq. (8).
    end for
  until convergence
  (M-Step:)
  update φ_vk via Eq. (12) and Eq. (13)
  update β via Eq. (14)
  update π via Eq. (15) by Newton–Raphson algorithm.
end for
```

5. Sentence Representation Learning Algorithm with bi-DCSR

In the bi-DCSR method, we are interested in the latent sentence-topic portions ϑ^{s_j} that are enhanced by the two-directional contexts and the inner words as shown in Eq. (3). During the process of applying the variational EM algorithm

for the HPTM, we can obtain the topic proportions of $\theta^{s_j^i}, \theta^{s_{j-}^i}$, and $\theta^{s_{j-}^i}$. Here, we let $\theta^{s_j^i}$ denote the topic distribution of sentence $\mathbf{s}_j^i$ learned from the HPTM, which is the agent of $p(\vartheta^{s_j^i}|\theta^{s_{j-}}, \theta^{s_{j+}}, w_{j1}^i, \cdots, w_{jN_j^i}^i)$.

We need to update the topics of $\mathbf{s}_j^i$ by the bi-DCSR method using Eq. (3).

For topic z_k, we compute the probability of z being assigned to the words in sentence $\mathbf{s}_j^i$ as

$$p(z_k|w_{j1}^i, w_{j2}^i, \cdots, w_{jN_j^i}^i) = \frac{p(z_k|\vartheta^{s_j^i}, w_{j1}^i, w_{j2}^i, \cdots, w_{jN_j^i}^i)}{\sum_{z\prime\in T} p(z\prime|\vartheta^{s_j^i}, w_{j1}^i, w_{j2}^i, \cdots, w_{jN_j^i}^i)}, \tag{16}$$

where we have

$$p(z_k|\vartheta^{s_j^i}, w_{j1}^i, w_{j2}^i, \cdots, w_{jN_j^i}^i) = p(\vartheta^{s_j^i}|w_{j1}^i, \cdots, w_{jN_j^i}^i; z_k).$$

By combining Eqs. (3) and (16), the expanded proportion of z_k can be computed as

$$p(z_k|w_{j1}^i, w_{j2}^i, \cdots, w_{jN_j^i}^i) = \frac{\theta_{z_k}^{s_j} \cdot \theta_{z_k}^{s_{j-}} \cdot \theta_{z_k}^{s_{j+}} \cdot \prod_{n=1}^{N_j^i} \phi_{z_k}^{w_{jn}^i} \cdot \sum_{n=1}^{N_j^i} \beta_{z_k}^{w_{jn}^i}}{\sum_{z\prime\in T} \vartheta_{z\prime}^{s_j} \cdot \theta_{z\prime}^{s_{j-}} \cdot \theta_{z\prime}^{s_{j+}} \cdot \prod_{n=1}^{N_j^i} \phi_{z\prime}^{w_{jn}^i} \cdot \sum_{n=1}^{N_j^i} \beta_{z\prime}^{w_{jn}^i}}. \tag{17}$$

All of the values in Eq. (17) are computed by the HPTM and we update $\vartheta^{s_j^i}$ in bi-DCSR. The updated algorithm of the bi-DCSR method is shown in Algorithm 3.

Algorithm 3: Update process of bi-DCSR.

1: INPUT: Sentences $(\mathbf{s}_1^i, \mathbf{s}_2^i, \cdots, \mathbf{s}_{S^i}^i)$ in a corpus C, ϕ, and β learned by Algorithm 2.
2: OUTPUT: The topic distributions of the sentences $\vartheta^{s_j^i}$ by bi-DCSR.
3: Set the contextual window size C_r.
4: Inference HPTM to obtain θ^s for all the sentences.
5: **for**Each sentence $\mathbf{s}_j^i$ of each document $\mathbf{d}^i$ in **bdo**
6: Get the topic distributions of the preceding sentences $\theta^{s_{j-C_r}}$, and the subsequent sentences $\theta^{s_{j+C_r}}$.
7: Get the topic distribution of sentence $\theta^{s_j^i}$.
8: **for** Each topic z_k, **do**
9: Compute $p(z_k)$ via Eq. (17).
10: **end for**
11: **end for**

6. Complexity analysis

We discuss the time complexity of the variational inference for the HPTM and the bi-DCSR method, compared with the traditional mean-field variational inference of the LDA method. For clarification, we again declare that K is the number of topics, M denotes the number of documents in the corpus, N_d is the number of words in a document, S is the number of sentences in a document, N_s is the number of words in a sentence, V denotes the size of the dictionary, and C_r is the contextual window size used in the bi-DCSR method.

For the LDA method, the mean-field variational inference requires, through all of the documents, to compute the variational parameters over K topics per iteration. Thus, the per-iteration time complexity of the LDA method is $O(M \times N_d \times K)$. Similarly, that of the HPTM is $O(M \times S \times N_s \times K)$. Since $N_d \gg N_s$ and $N_d \approx S \times N_s$, the time complexity of the HPTM is close to the LDA method per-iteration.

From Algorithm 3, we can see that the bi-DCSR methods needs to inference the HPTM through the whole corpus first, and then run the context-enhanced process through the target sentences. Here, we assume all of the sentences in the corpus are involved in the context-enhanced process. Thus, the time complexity of the bi-DCSR methods is $O(M \times S \times N_s \times K + M \times S \times K \times (N_s + C_r))$ in the worst case. Actually, the HPTM needs iterations to achieve model convergence; thus, the time complexity of the bi-DCSR method is much smaller than the HPTM. Because the bi-DCSR method depends on the outputs of the HPTM, the online algorithm of the HPTM greatly reduces the complexity of the bi-DCSR method in many real applications. In general, the time complexity of both algorithms is dominated by the cost of the embed-

ding process of the sentences and words with the same topic space in the HPTM. This is highly dependent on the scale of the corpus, but the online algorithm can handle the worst case.

7. Experiments

In this section, we will evaluate the proposed method on document modeling, sentence classification tasks, and semantic interpretability. The codes and data used in the experiments are released on an open website[3].

7.1. Datasets

To prove the concept of the proposed model and investigate its performance, four different datasets have been chosen from open and closed domains. The datasets used are described as follows.

(1) Wikipedia. Wikipedia is a typical open domain dataset. The abstract of each page is extracted from the Wikipedia dump and remove the pages with less than 10 sentences. This subset of Wikipedia contains 160,942 pages. 34,956 pages are selected for testing the HPTM, and the remaining 125,986 pages are for training the HPTM. The testing pages belong to 48 categories, such as education, book, history, military, etc., which are class labels for sentence classification tasks. Each page in the test set belongs to one of the categories. We remove stop words and words which appear less than 10 times.
(2) arXiv. This dataset is from arXiv.org, which is a collection of bibliographic proceedings. We extract the abstracts of papers from arXiv[4] between 2018–01-01 and 2019–04-01, and obtain 216,146 abstracts. We select 23,815 abstracts from arXiv for testing the HPTM, and the remaining 19,2331 pages are for training. The testing pages belong to nine categories, such as computer science, mathematics, physics, statistics, and so on. We remove stop words and words which appear less than 5 times.
(3) LAW. This dataset is from CAIL[5], which is a competition on document classification tasks of court verdicts. A subset of the CAIL are extracted, which contains the judgment documents on financial fraud, such as fraudulent credit-card, fraudulent fund-raising, contract fraud, insurance fraud, and so on. There are six types of judgment documents on financial fraud, and the total number is 4,060. 3,060 judgment documents as the training set are selected, and the remaining 1,000 are for testing. We use jieba[6] as the Chinese text segmentation module, and keep the top 3,000 words as the dictionary.
(4) MedicalTS. This dataset is from Collection of Transcribed Medical Transcription Sample Reports and Examples[7], which contains sample transcription reports for many specialties and different types of work. There are 5,000 transcription texts with 40 different specialties in total. We use the specialties as the label information. We remove the specialties with very little transcription texts, which results in 3,104 samples within nine specialties, such as cardiovascular/ pulmonary, neurology, radiology, urology, and so on. From this set, we randomly select 2,500 samples for training; the remaining 604 are for testing.

Table 2 highlights the summary statistics of the datasets.

7.2. Language Modeling

This subsection demonstrates the performance of the proposed HPTM on language modeling. The evaluation matrices include perplexity and topic coherence.

7.2.1. Analysis of the HPTM

In this part of the experiments, we analyze the proposed HPTM on the contextual window size C and the topic dimensionality K. The contextual window size C controls how many contextual sentences are used for the sentence representation learning. We evaluate a range of values for C on the four corpora. For each corpus, we run the HPTM on the training set, and use the held-out perplexity on the test data as a measure of the model fit to show the effects of contextual window size. Meanwhile, to analyze the parametric sensitivity with the dimensionality of sentence embeddings, we test different numbers of topics K, which are the dimensions of the sentence representation learned by our method.

The perplexity is commonly exploited in measuring the capabilities of topic models. The perplexity on the sentence level is defined as:

[3] http://www.shuangyin.li/bidcsr.html
[4] http://export.arxiv.org
[5] http://cail.cipsc.org.cn
[6] https://github.com/fxsjy/jieba
[7] https://www.mtsamples.com

Table 2
Summary statistics of Wikipedia, arXiv, IMDB, and MedicalTS.

#	Wikipedia	arXiv	LAW	MedicalTS
# of documents	160,942	216,146	4,060	3,104
# of sentences	1,679,866	1,355,068	34,582	119,580
# of words	12,853,000	14,231,362	689,088	731,110
length of the dictionary	20,967	18,222	3,000	18,009
mean # of words per sentence	7.65	10.5	19.9	6.11
mean # of sentences per document	10.44	6.27	8.5	38.52

$$Perplexity(D_{test}) = \exp(-\frac{\sum_{i=1}^{M_{test}}\sum_{j=1}^{S_i}\log p(\mathbf{s}_j^i)}{\sum_{i=1}^{M_{test}}\sum_{j=1}^{S_i}N_j^i}), \tag{18}$$

where, M_{test} represents the number of documents in the test set. The lower perplexity value on the test set is, the better performance of the language modeling is. Since the value of $\log p(\mathbf{s}_j^i)$ cannot be computed directly, the Jensen's lower bound on the log probability of the sentence $\mathbf{s}_j^i$ is used as a proxy defined in Eq. (6). In this part of the experiments, the number of topics is set to $K = 200$.

According to the average number of sentences per document in different datasets, we set $C = 1, 2, 3, 4, 5$ for Wikipedia, arXiv, and LAW, and set $C = 1, 4, 8, 12, 16$ for MedicalTS. $C = 2$ means that two preceding sentences and two subsequent sentences are considered as the contexts. For the sentences in the head or tail of a document, e.g., $i < C$, we treat the sentences themselves as their contextual sentences. For the topic dimensionality K, we use the following settings for each corpus: $K_{Wikipedia} = 50, 100, 200, 300, 500, K_{arXiv} = 50, 100, 150, 200, 300, K_{LAW} = 10, 50, 100, 150, 200,$ and $K_{MedicalTS} = 10, 50, 100, 150, 200$.

We examine the held-out perplexities on the test sets. Held-out likelihood metrics are well suited to measure a topic model. Fig. 5 summarizes the results for each corpus. According to the results, we find that the best number of contexts for the four corpora is different, e.g., $C = 3$ for Wikipedia, $C = 5$ for arXiv, $C = 3$ for LAW, and $C = 4$ for MedicalTS. In addition, we obtain the best K for each corpus, e.g., $K = 500$ for Wikipedia, $K = 200$ for arXiv, $K = 50$ for LAW, and $K = 50$ for MedicalTS. Large values of K incur model over-fitting, especially on LAW and MedicalTS. In the following experiments, we use the best settings of C and K for each corpus.

In order to test the online training algorithm, we also design the following experiments. Since the datasets of LAW and MedicalTS are small, Wikipedia and arXiv are used in this part of experiments. Based on the online training algorithm of the HPTM, we split the training sets from Wikipedia and arXiv into several mini-batches. Each mini-batch contains 2000 documents. The online parameter settings for the two corpora are $\tau_0 = 64$ and $\eta = 0.5$ with $K = 500$ and $K = 200$, respectively. Fig. 6 demonstrates the results. We find that the perplexities are consistently reduced as the number of training batches increases, which verifies the convergence of the online algorithm.

7.2.2. Performance on Perplexity and Topic Coherence

As the HPTM is the core algorithm for learning the sentence representation, we compare the HPTM with five other topic models on held-out perplexity and topic coherence metrics for the purpose of investigating and demonstrating the performance on language modeling. To quantitatively analyze the quality of the topics, we use two different coherence measures, the UCI measure and the UMass measure, to help distinguish between the topics learned by the HPTM. The topic coherence measures the degree of semantic similarity among the top words in a topic [43,44].

The UCI measure for topic k is defined as

$$UCI(w_i, w_j)_k = \log\frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{19}$$

where, $p(w)$ indicates the probability of w in a document, and $p(w_i, w_j)$ is the probability for the two words co-occurring in one document. A higher UCI score indicates a more semantically coherent topic [44].

The UMass measure for topic k is defined as

$$UMass(w_i, w_j)_k = \log\frac{D_{test}(w_i, w_j) + 1}{D_{test}(w_i)}, \tag{20}$$

where, $D_{test}(w_i)$ denotes the counts of w_i in the test set. Similar to the UCI measure, the higher the UMass measure is, the better the topic is, where the words with high semantic similarity are clustered. For the UCI and UMass measures, we rank the words for each topic by their probabilities, and then compute the scores within the top 20 words on the test dataset.

For comparative study, we compare the proposed approach with the following topic models:

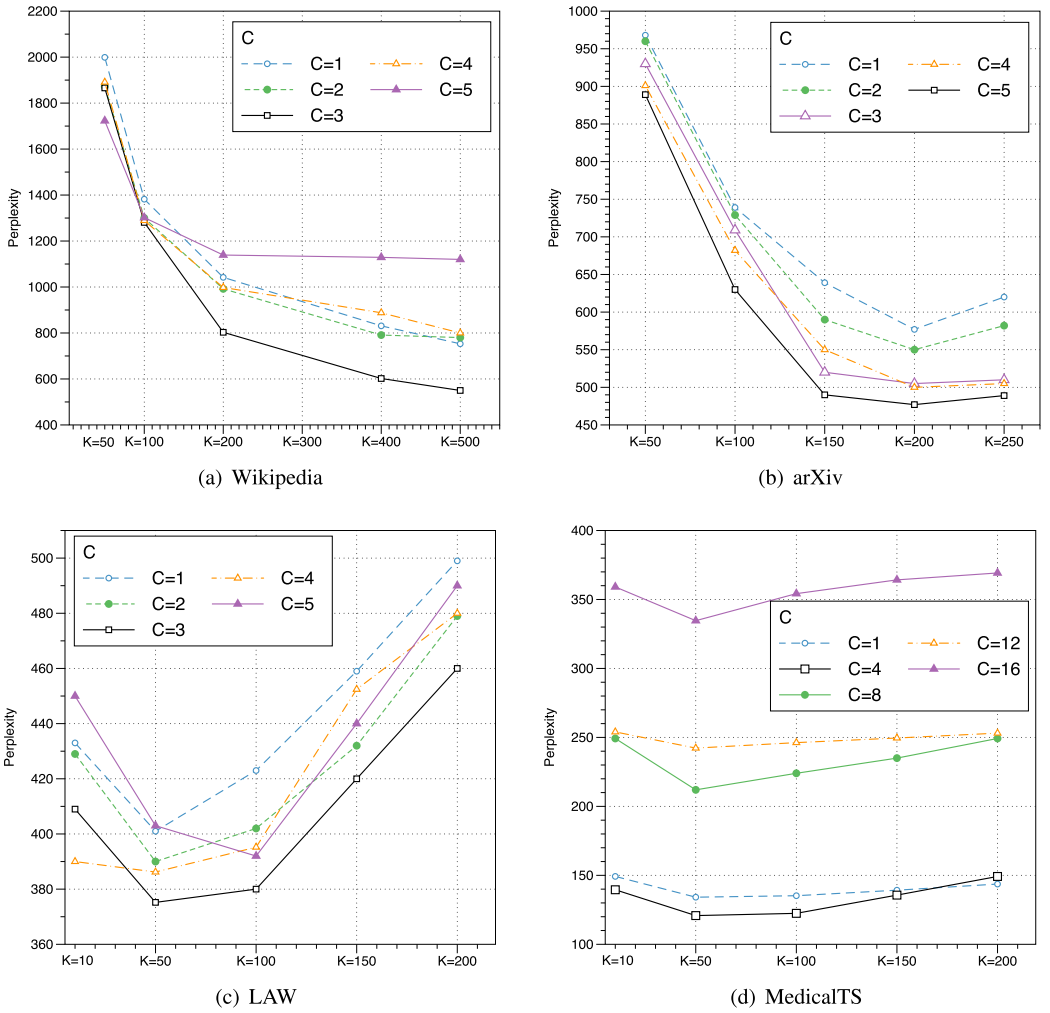


Fig. 5. The held-out perplexities obtained on the test sets of Wikipedia, arXiv, LAW and MedicalTS with different values of C and K. The x-axes denote the topic numbers.

(1) LDA. We train the LDA model on the sentence level with the implementation from the standard toolkit.[8] The em_max—_iter is 100, and the em_convergence is 1e-4. We set the alpha to be estimated and gradually reduced to the final values, for example, $\alpha_{Wikipedia} = 0.0012, \alpha_{arXiv} = 0.001, \alpha_{LAW} = 0.0105$ and $\alpha_{MedicalTS} = 0.004$.

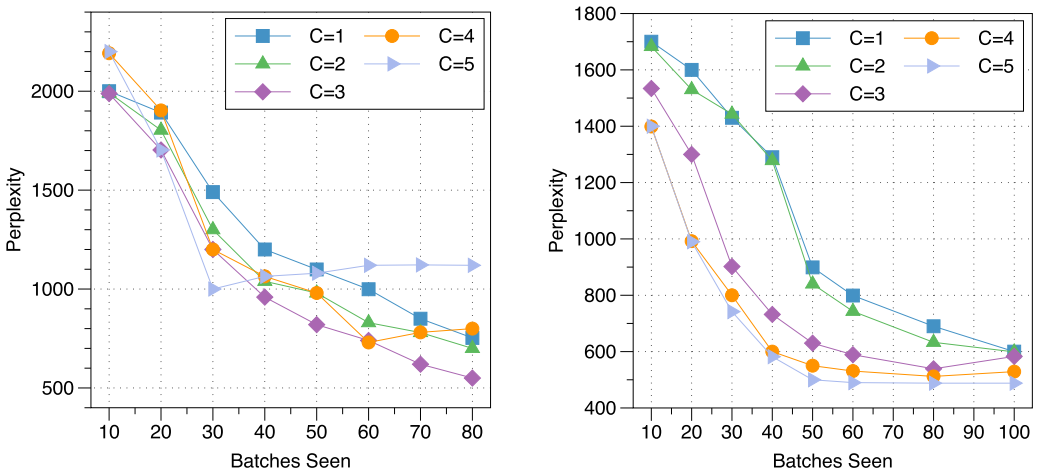
(2) Hierarchical Dirichlet processes (HDP). HDP is an unsupervised topic model that handles a corpus without considering the initial number of topics. We use the implementation from a public toolkit[9]. The training parameters for the four corpora are set as follows, $gamma_a$ = 1.0, $gamma_b$ = 1.0, $alpha_a$ = 1.0, $alpha_b$ =1.0 and eta = 0.5.

(3) Gaussian Mixture Neural Topic Model (GMNTM) [45]. GMNTM learns the topic model and the document representations, where the words' sequential information are considered to model the semantics of the document. For the four corpora, the *alpha* in GMNTM is set to 0.025 and m is set to 6.

(4) Sequential Latent Dirichlet Allocation (SeqLDA) [46]. SeqLDA considers the segments of a document, e.g., the chapters and paragraphs, as the underlying sequential structure. In our experiments, we consider the sentences as the segments of a document. The SeqLDA is trained following the guide in [46] to search the optimised parameters, where $\alpha = 0.1$, a = 0.2, and b is optimized for each segment.

(5) Recurrent Attentional Topic Model (RATM) [13]. RATM models the documents as a group of sequences of sentences, where only the preceding sentences are involved. We use the released implementation. For a fair comparison, The same number of preceding sentences is considered as the contexts for both RATM and our model. Differently, our model considers more following sentences of the target sentence as the contexts. For the four corpora, the window in RATM is set to 4, the em_max—_iter is 100, and the em_convergence is 1e-6.

[8] http://www.cs.columbia.edu/~blei/lda-c/index.html
[9] https://github.com/blei-lab/hdp



(a) Wikipedia (b) arXiv

Fig. 6. The held-out perplexities on Wikipedia and arXiv using the online algorithm. The perplexities are computed using the model parameters learned with the continually seen batches.

(6) DocNADE [47]. The DocNADE is an undirected graphical model of word counts that was shown to learn a better generative model and more meaningful document representations based on replicated Softmax. We use the released implementation[10], and the default parameters are used.
(7) MetaLDA [48]. The MetaLDA handles the sparse texts, such as the sentences with less word information, by the meta information. The released implementation[11] is sorted to compute the perplexity and topic coherence values. The α and β are both set to 1.0 for each corpus.

In this part of the experiments, we choose GMNTM, SeqLDA, and RATM for comparison, because they all consider the sequential structures within different levels, not only the word information just like LDA and HDP. Such as RATM models the documents as a group of sequences of sentences, where only the preceding sentences are involved. We use the released implementation[12]. For a fair and objective comparison, the same number of preceding sentences is considered as the contexts for both RATM and our model. Differently, however, our model considers the extra subsequent sentences of the target sentence as the contexts. Moreover, the DocNADE and the MetaLDA are also considered as comparisons that DocNADE is one of the generative topic models with neural network architecture, and the MetaLDA is to handle the texts with less word information, such as the sentences.

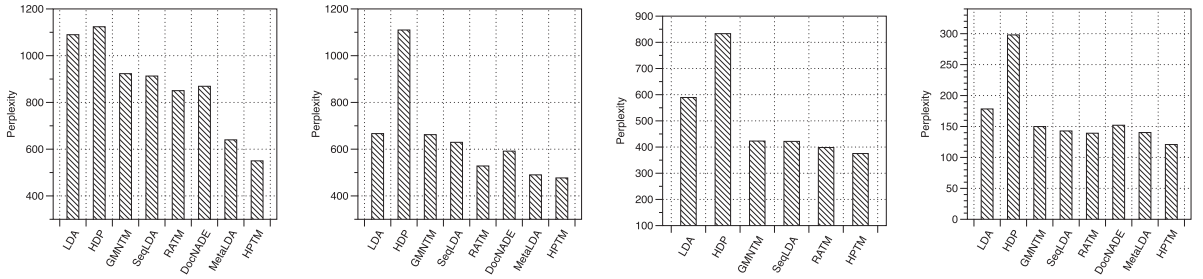
The held-out perplexities, UCI measures, and UMass measures are computed on the test sets. Fig. 7 compares the held-out perplexities on Wikipedia, arXiv, LAW, and MedicalTS. Our proposed model takes advantage of more information than LDA and RATM on sentence modeling, which is why our model achieves better results. For example, RATM takes advantage of the information of only the preceding sentences for topic extraction, while, our model considers the bi-directional contexts and the inner words. Compared with the MetaLDA, the HPTM gets benefit from the plentiful word and contextual information. According to the results, the proposed HPTM outperforms other topic models in the experiments.

Table 3 reports the UCI and UMass scores of LDA, GMNTM, SeqLDA, RATM, and the proposed HPTM. We compute the UCI and UMass scores by considering the top 20 words in each topic, where the words in each topic are ranked by their probabilities. We first compute the scores for each topic, and the average scores are shown for all of the topics. The proposed HPTM achieves higher average scores on Wikipedia, LAW, and MedicalTS. The documents in arXiv are usually short and have a narrow topic-focus, so LDA outperforms the other baselines for arXiv. In this case, however, the HPTM still exhibits a competitive performance. For the Wikipedia, the MetaLDA obtains the better results on UCI than the HPTM, while, the HPTM achieves the better performances on the closed domains in LAW and MedicalTS. The performance on perplexity and topic coherence indicate the effectiveness of the proposed HPTM.

7.3. Evaluation on Sentence Classification

The main contribution of this work is the proposal of a framework of sentence representation learning in closed-domains; thus, we evaluate our method on sentence classification tasks. With the HPTM, we obtain ϕ, which is the topic distribution of

[10] https://github.com/pgcool/iDocNADEe
[11] https://github.com/ethanhezhao/MetaLDA
[12] https://github.com/shuangyinli/RATM



(a) Wikipedia (b) arXiv (c) LAW (d) MedicalTS

Fig. 7. The held-out perplexities of different models on the test sets.

Table 3
The results of UCI and UMass scores on Wikipedia, arXiv, LAW and MedicalTS.

	UCI				UMass			
	Wikipedia	arXiv	LAW	MedicalTS	Wikipedia	arXiv	LAW	MedicalTS
LDA	228.46	452.33	36.4	213.21	−1011.89	−395.13	−2889.33	−1239.61
GMNTM	198.35	329.81	30.10	192.09	−1230.55	−484.65	−2933.09	−1145.50
SeqLDA	200.97	333.56	25.64	189.24	−1331.50	−540.88	−2900.12	−1031.01
RATM	229.01	430.98	36.66	219.66	−1009.86	−399.11	−2895.79	−1003.62
MetaLDA	231.22	450.19	36.79	215.01	−972.35	−396.34	- 2822.25	- 999.03
HPTM	230.21	447.18	37.91	222.89	−937.94	−396.03	−2814.22	−979.77

the words in the dictionary, and β, which is the word probabilities of the latent topics. With the formulation of sentence embedding shown in the bi-DCSR method, we can learn the context-enhanced representations for the sentences in the test sets based on the output of the HPTM with Algorithm 3.

Different from existing methods that focus on the syntactic aspects of a sentence [49], we are more interested in assessing how well the representations capture the sentence semantics in some closed-domains. We construct multi-class classification tasks using the one-against-all strategy for the sentence classification tasks by assigning the sentences with the label of the host document. For LAW and MedicalTS, a fivefold cross-validation is used for quality classification (four folds for training and one fold for testing).

We compare the proposed approach with the following state-of-the-art unsupervised methods. We summarize the parameters of each comparison in Table 4.

(1) Word2Vec (BoW). We train Word2Vec(Skip-gram) using the implementation of Gensim [13] on the four datasets, where we set the different dimensions of the word vectors for each corpus, such as 500 on Wikipedia.
(2) FastText (BoW). This code is obtained from the website[14]. We train the FastText to obtain the 500-dimensional word vectors. Like Word2Vec, a traditional bag-of-words averaging is employed to produce the sentence embedding.
(3) ELMo (BoW). ELMo is a deep contextualized word representation learning method. We use the released implementation[15] and retrain ELMo models on our datasets. Moreover, we compare the pre-trained ELMo with the default parameters, which was trained on a dataset of 5.5 B tokens consisting of Wikipedia (1.9 B) and all of the monolingual news crawl data from WMT 2008–2012 (3.6 B). Also, a bag-of-words averaging is employed to produce the sentence embedding. Note thatthe vectors of the words that do not appear in the pre-trained ELMo, e.g., entities in MedicalTS, are set to the value 0.
(4) AutoEncoder. We implement an AutoEncoder with the sentences as the input to obtain the sentence vectors. The AutoEncoder embeds the sentences into a 500-dimension space. Since it is difficult for an AutoEncoder to converge on small datasets, i.e., LAW and MedicalTS, we train the model on Wikipedia and arXiv.
(5) Paragraph2Vec. We use the implementation from Gensim, and Paragraph2Vec learns paragraph and document embeddings via the distributed memory and distributed bag of words models.

[13] https://radimrehurek.com/gensim/
[14] https://fasttext.cc
[15] https://allennlp.org/elmo

Table 4
Summary of the dimensions of the sentences' vectors using by the comparisons on Wikipedia, arXiv, IMDB, and MedicalTS.

#	Wikipedia	arXiv	LAW	MedicalTS
Word2Vec (BoW)	500	200	50	50
FastText (BoW)	500	200	50	50
ELMo (BoW)	500	200	-	-
Pre-trained ELMo (BoW)	512	512	512	512
AutoEncoder	500	200	-	-
Paragraph2Vec	500	200	50	50
Skip-Thought	500	200	-	-
Quick-Thought	500	200	-	-
Pre-trained Bert	768	768	768	768
ClinicalBert	-	-	-	768
LDA	500	200	50	50
RATM	500	200	50	50
HPTM	500	200	50	50
bi-DCSR	500	200	50	50

(6) Skip-Thought. The model is obtained from the authors' website[16]. In our paper, we use the code provided by the authors and train bi-skip models with our corpora. For example, on Wikipedia we train a bidirectional model with forward and backward encoders of 250 dimensions each. Then, we concatenate the two vectors from the forward and backward encoders, resulting in 500-dimension sentence vectors.
(7) Quick-Thought. We use the implementation of Quick-Thought from the website[17]. Here, we retrain the Quick-Thought to obtain 500-dimension sentence vectors on Wikipedia and 200-dimension sentence vectors on arXiv.
(8) Bert. We use the pre-trained Bert, which is trained on Wikipedia and BookCorpus[18]. To obtain the representations of the sentences, we use the implementation of bert-as-service[19], which uses Bert as a sentence encoder and hosts it as a service. For LAW, the Chinese-Bert is adopted. All of the models are fine-tuned with the training sets. We use the output of the special [CLS] token in Bert as the sentence vectors. Also, the vectors of the words that do not appear in the pre-trained Bert are set to the value 0.
(9) ClinicalBert. The ClinicalBert is pre-trained on clinical notes to output a binary prediction of whether these two sentences are in consecutive order using the released implementation[20]. The vectors of the words that do not appear in the pre-trained ClinicalBert are set to the value 0 as the Bert does.

For our proposed model, we first train the HPTM on the given corpora taking the sequences of the sentences as input. For example, we set the number of the topics to be $K = 500$ on Wikipedia, which means the dimension of the sentence representation is 500. We treat the sentence representations learned by the HPTM as one of the comparisons. After that, we run Algorithm 3 to obtain the context-enhanced sentence features by the bi-DCSR method. We set $C = C_r$, which means the bi-DCSR method uses the same number of neighbored sentences as the HPTM.

In addition, we compare the models which are based on the framework of topic modeling. We train the LDA and RATM at the sentence level. Since RATM considers the preceding sentences, we set the number of preceding sentences to be the same as for our model.

As a context-enhanced sentence representation learning framework, the proposed bi-DCSR method can also take advantage of LDA to obtain context-enhance sentence embeddings. The procedure is as follows. We first train LDA at the sentence level, and inference on the test set to obtain the topic distributions for each sentence. Then, we run Algorithm 3 with the preceding and subsequent neighbored sentences as the contexts. Different from the bi-DCSR method, the LDA-bi-DCSR does not consider the topics of the inner words, as LDA cannot learn the words' topic distributions.

Likewise, we run RATM to obtain the sentence embeddings, and run Algorithm 3 to update the context-enhanced sentence representation by the bi-DCSR method, which is called RATM-bi-DCSR. Also, we implement two variants of the bi-DCSR method. (1) We only utilize the inner words as the contexts of the sentence, and compute the sentence features by Algorithm 3. We call this variant bi-DCSR-W. (2) We only consider the neighboring sentences as the contexts; we call this variant bi-DCSR-S. We use these two variants to test the impact of neighboring sentences.

We use a Support Vector Machine(libSVM)[21] with a Gaussian kernel as the classifier, and test the accuracy and F1-score for all of the models. The experimental results of the proposed method and other comparisons on sentence classification tasks are presented in Table 5.

[16] https://github.com/ryankiros/skip-thoughts
[17] https://github.com/lajanugen/S2V
[18] https://github.com/google-research/bert
[19] https://github.com/hanxiao/bert-as-service
[20] https://github.com/kexinhuang12345/clinicalBERT
[21] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Table 5
Sentence classification results of unsupervised models on Wikipedia, arXiv, LAW and MedicalTS. Accuracy and F1 are presented. For LAW and MedicalTS, a fivefold cross-validation is used for quality classification.

Models	Wikipedia		arXiv		LAW		MedicalTS	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Word2Vec	0.449	0.433	0.622	0.559	0.630(±0.008)	0.615(±0.007)	0.582(±0.010)	0.571(±0.009)
FastText	0.444	0.430	0.612	0.549	0.612(±0.006)	0.601(±0.005)	0.577(±0.005)	0.543(±0.004)
ELMo	0.459	0.454	0.633	0.610	-	-	-	-
Pre-trained ELMo	0.468	0.457	0.641	0.592	0.661(±0.008)	0.639(±0.007)	0.521(±0.009)	0.506(±0.011)
AutoEncoder	0.267	0.243	0.405	0.349	-	-	-	-
Paragraph2Vec	0.093	0.016	0.330	0.163	0.181(±0.004)	0.129(±0.003)	0.110(±0.002)	0.089(±0.003)
Skip-Thought	0.347	0.280	0.590	0.470	-	-	-	-
Quick-Thought	0.385	0.326	0.589	0.490	-	-	-	-
Pre-trained Bert	0.480	0.472	0.631	0.599	0.711(±0.004)	0.696(±0.003)	0.567(±0.005)	0.533(±0.004)
ClinicalBert	-	-	-	-	-	-	0.662 (±0.004)	0.641 (±0.003)
LDA	0.297	0.256	0.501	0.439	0.582(±0.001)	0.521(±0.002)	0.566(±0.002)	0.508(±0.002)
RATM	0.363	0.352	0.529	0.464	0.699(±0.002)	0.630(±0.002)	0.639(±0.003)	0.600(±0.003)
HPTM	0.423	0.413	0.593	0.555	0.719(±0.003)	0.688(±0.002)	0.648(±0.004)	0.625(±0.003)
LDA-bi-DCSR	0.392	0.357	0.598	0.551	0.628(±0.002)	0.600(±0.002)	0.573(±0.002)	0.509(±0.002)
RATM-bi-DCSR	0.424	0.409	0.601	0.559	0.704(±0.003)	0.666(±0.004)	0.643(±0.003)	0.611(±0.005)
bi-DCSR-W	0.453	0.443	0.623	0.565	0.733(±0.003)	0.701(±0.003)	0.656(±0.002)	0.631(±0.003)
bi-DCSR-S	0.432	0.429	0.618	0.563	0.720(±0.004)	0.704(±0.003)	0.650(±0.003)	0.609(±0.002)
bi-DCSR	0.460	0.456	0.634	0.575	0.739(±0.002)	0.710(±0.003)	0.673(±0.003)	0.655(±0.002)

Table 5 shows that, bi-DCSR yields better performance compared to other approaches on arXiv, LAW, and MedicalTS, which corroborate the effectiveness of the proposed method, especially on these closed domains. To compare the unsupervised learning methods for sentence representations such as Skip-Thought, Quick-Thought, and AutoEncoder, bi-DCSR achieves better performances than that of all the above mentioned methods. The main reason may be that the deep learning methods need large-scale training data for model fitting. When it comes to small datasets, the performance of these models are worse than the models that can be trained from scratch. The pre-trained ELMo and the pre-trained Bert outperform the other models on Wikipedia, since the pre-trained models are well-fit on the large-scale Wikipedia. For these cases, however, bi-DCSR is still competitive. Table 5 shows that, the bi-DCSR yields better performance compared to the other approaches on LAW and MedicalTS, which corroborates the effectiveness of the proposed method, especially on these closed domains.

It is observed that the pre-trained ELMo and the pre-trained Bert achieve better results on arXiv, because the datasets used in the pre-training of the two models overlap with arXiv to some extent. Such as, pre-trained ELMo was trained on a dataset of 5.5B tokens consisting of Wikipedia (1.9B) and all of the monolingual news crawl data from WMT 2008–2012 (3.6B). Also, we find the trends of the results of the pre-trained models from open domain to closed domain are that our methods significantly outperform these pre-trained algorithms on closed domains.

The ClinicalBert is designed for the scenarios of medicine, thus, we show the results of sentence classification on MedicalTS. We can observe that the ClinicalBert obtains the better results than the pre-trained Bert, since that it is trained on the medical corpus. Even that, the bi-DCSR yields better performances than the ClinicalBert because the bi-DCSR is well-trained in such a closed domain with the limited data.

In Table 5, the "-" symbol denotes that we cannot train the model well because of a lack of data, such as ELMo, AutoEncoder, Skip-Thought, and Quick-Thought, if there is no pre-trained model. This is a drawback of the deep learning or transformer-based methods. Even if we can train them on other large scale datasets and then transfer them to the closed domains, they may still perform badly. This is because many new elements are not trained, such as special entities and proper nouns. In contrast, an important benefit of our method is that it can train on small datasets from scratch, which is useful in many special scenarios.

As one type of Bayesian generating model, the HPTM obtains better performance than that of the LDA approach. As described in [15], the length of documents plays a crucial role: poor performance of the LDA is expected when documents are too short. The proposed HPTM overcomes this drawback by considering the context information for the sentence-level modeling.

We note that LDA-bi-DCSR achieves better results than the LDA approach, which shows that the proposed context-enhanced mechanism works on sentence representation learning. Furthermore, bi-DCSR has better performance than LDA-bi-DCSR, as the bi-DCSR embeds the sentences using both sentence contexts and inner words. In addition, bi-DCSR-W outperforms bi-DCSR-S, which indicates the inner words have a great effect on sentence embedding. The semantics of the inner words in a sentence plays an important role in embedding sentences, supported by the evidence of the good performance of ELMo in Table 5. The contributions of using words and contexts' semantics are from the flexible modeling of the HPTM and the topic-enhanced algorithm of bi-DCSR, which are beneficial for sentence embedding.

7.4. Evaluation on Sentence Similarity

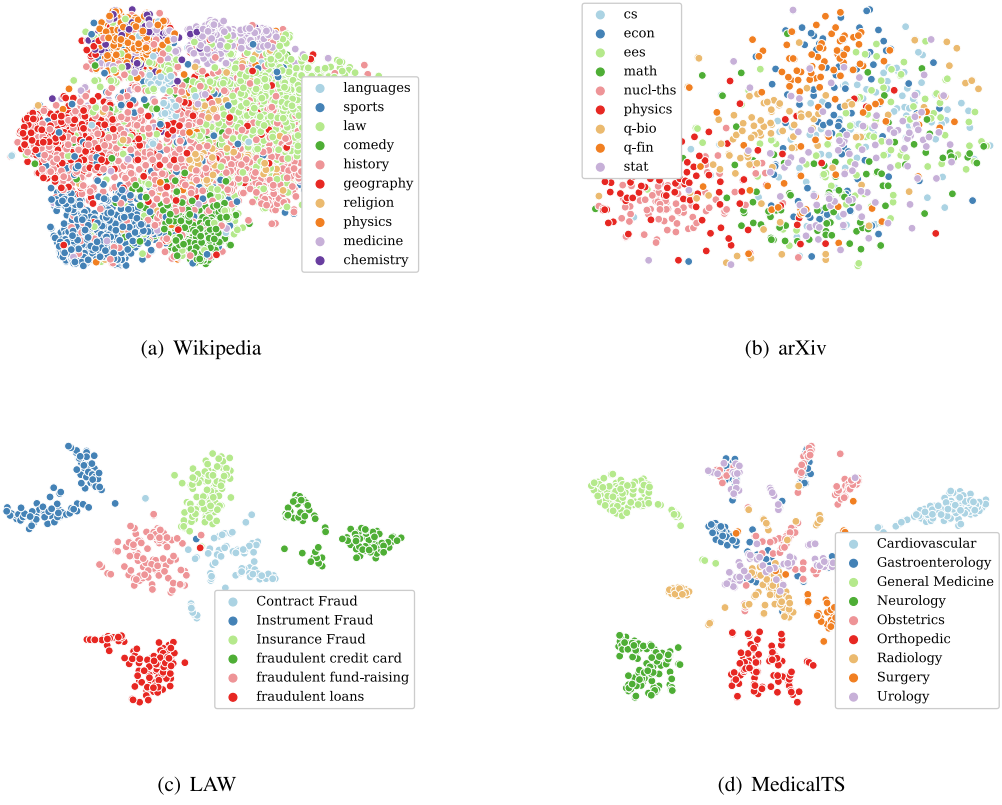
In this section, we evaluate the proposed model on the task of sentence similarity. Two test data sets with sentence pairs are conducted from Wikipedia and MedicalTS, and this is to predict the similarity for the given sentence pair. For Wikipedia, 10,000 sentence pairs are selected. Each pair takes the form of assigning a score by 0 or 1, where 1 indicates the two sentences are from the same abstract. Similar, 1,000 sentence pairs from MedicalTS are selected. The Pearson correlation is used to report the performance, and the higher Pearson correlations indicate better results.

Table 6 summaries the results on the task of sentence similarity. The Pre-trained ELMo, Pre-trained Bert, ClinicalBert, Paragraph2Vec and LDA are chosen as the comparisons, which are trained following the above settings. The performance of the bi-DCSR on the MedicalTS outperforms the other comparisons, which indicates that our method can effectively capture the semantics of the sentences in the closed domain. Even though the Pre-trained ELMo and Pre-trained Bert obtain the higher Pearson correlations on the Wikipedia, the bi-DCSR can also get the competitive result on the open domain.

Meanwhile, to make it more intuitive for the illustration of the capacity of the proposed framework on sentence representation learning, we apply t-SNE to visualize the representations of the sentences in Wikipedia, arXiv, LAW, and MedicalTS that are extracted by the bi-DCSR method. The visualization is shown in Fig. 8, where each point represents a sentence and the points are colored based on their labels. It can be observed that the points in Fig. 8 are clustered with the same class label, which makes it easy to perform the sentence similarity and demonstrate the effectiveness of the proposed framework of sentence representation.

Table 6
The performances of sentence representation models on the task of semantic similarity with Wikipedia and MedicalTS. The Pearson correlation is used to report the performances.

Models	LDA	Pre-trained ELMo	Pre-trained Bert	ClinicalBert	Paragraph2Vec	bi-DCSR
Wikipedia	0.81	0.91	<u>0.92</u>	-	0.59	0.90
MedicalTS	0.73	0.69	0.72	0.82	0.45	<u>0.83</u>



(a) Wikipedia (b) arXiv

(c) LAW (d) MedicalTS

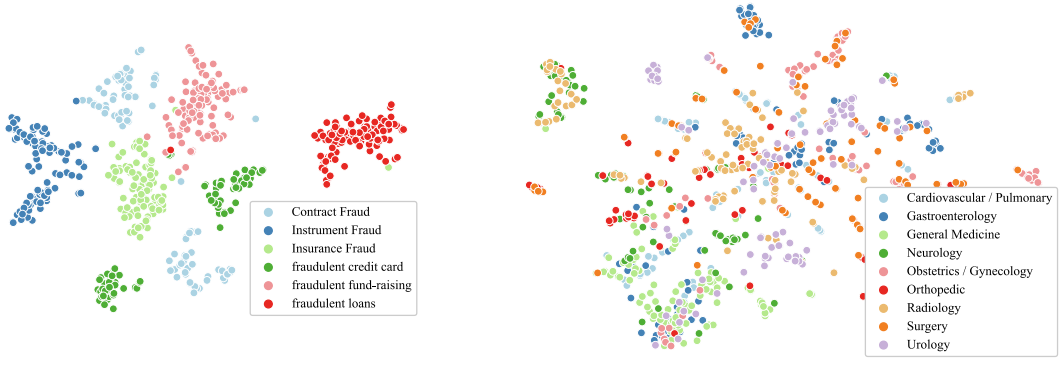
Fig. 8. The scatter plots of the selected categories on the Wikipedia, arXiv, LAW and MedicalTS, where a point denotes one sentence, and the class labels are shown in the legends.

To compare with other models, we also visualize the scatter plots of LDA and Paragraph2Vec. Since we need to demonstrate the effectiveness of on closed domains, we draw the scatter plots of LDA and Paragraph2Vec on LAW and MedicalTS. Fig. 9 show the results of LDA, and Fig. 10 show the results of Paragraph2Vec. On LAW, from Fig. 9 and Fig. 10, we find that LDA performs as well as the bi-DCSR method, and Paragraph2Vec performs worse. The main reason is that the neural-network-based methods, such as Paragraph2Vec, need more data to fit the models, while, the Bayesian methods, such as LDA and bi-DCSR have an edge on closed domains. Moreover, for the complicated domains, such as MedicalTS, the bi-DCSR works better than LDA and Paragraph2Vec, which shows the effectiveness of the context-enhanced process proposed in bi-DCSR.

7.5. Evaluation on Semantic Interpretability

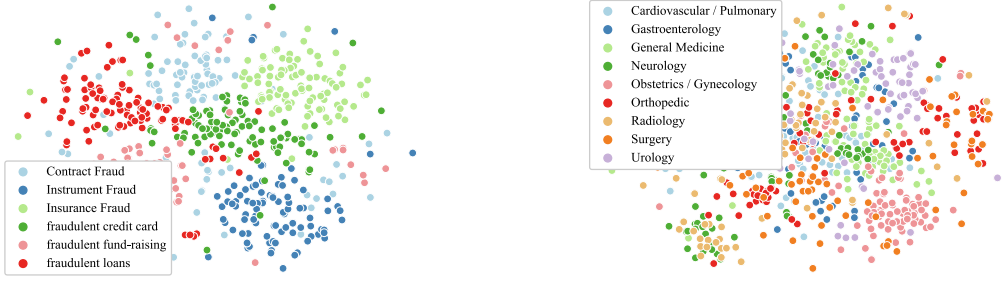
The proposed model aims to learn high-quality topic distributions of sentences and words, which suggests the examination of properties of the embedding space to better understand how it learns semantics. Thus, we design a nearest-neighbor retrieval experiment to test the capacity of the proposed method on capturing the sentence and word semantics. For a given query sentence, we rank the candidate sentences by the cosine distance in the embedding space, and the best neighbor is retrieved. We randomly select ten sentences from Wikipedia and arXiv, and retrieve the nearest sentence in the test sets for each of them. Table 7 shows the cases of query sentences from the dataset and the corresponding retrieved sentences. The results show that the query sentences are highly related to the retrieved sentences based on semantics. In addition, these results demonstrate the effectiveness of the context-enhanced learning process on the topic simplex, which is learned from the proposed approach.

Meanwhile, one benefit of the proposed framework is that the words and sentences are embedded into the same topic simplex, meaning that we can compute the similarity of a query sentence and a word. Thus, we show the top 10 nearest words for the selected sentences in Table 7, and visualize them by t-SNE in Fig. 11. This figure illustrates the capacity of the proposed framework; the bi-DCSR method is shown to be able to more effectively capture semantics through its bi-directional context-enhanced procedure than those of traditional methods.



(a) LAW (b) MedicalTS

Fig. 9. The scatter plots of LDA on the LAW and MedicalTS with the selected categories.



(a) LAW (b) MedicalTS

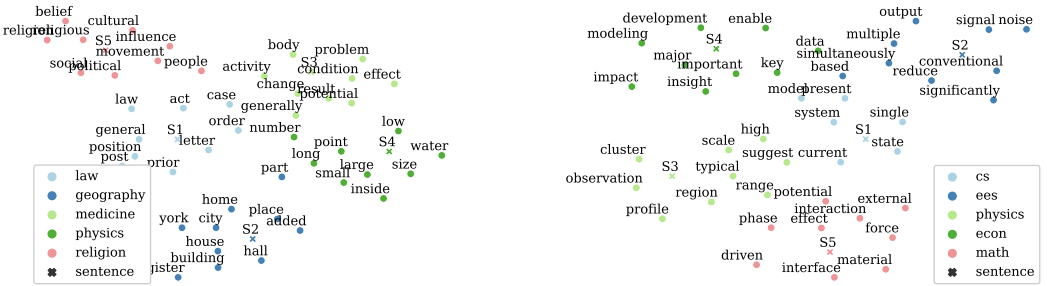
Fig. 10. The scatter plots of Paragraph2Vec on the LAW and MedicalTS with the selected categories.

Table 7
Some random cases of query sentences and the corresponding best neighbored sentence. The above five sentences are from Wikipedia, and the down five sentences are from arXiv.

S1: A public interest litigation cell has been opened in the Supreme Court to which letters addressed to the Court or individual judges are forwarded, which are placed before the Chief Justice after scrutiny by the staff attached to the cell.
Bestneighbor: In accordance with Article 145 of the Constitution of India and the Supreme Court Rules of Procedure of 1966, the Chief Justice allocates all work to the other judges who are bound to refer the matter back to him or her (for re-allocation) in any case, where they require it to be looked into by a larger bench of more judges.
S2: Prominent architecture located within the district include the Cobblestone Path Nelson County Jail Old L and N Station Old Talbott Tavern and Spalding Hall all individually on the National Register and the historic old Nelson County Courthouse.
Bestneighbor: Prominent public buildings on the common include the 1817 town hall, separately listed on the National Register the public library built as a private home in 1804, and sold to the town in 1912 and an early fire engine house.
S3: A statistical study published in 2005 tested potential risk factors: age, gender, body mass index, smoking, asthma, diabetes, cardiovascular disease, previous decompression illness, years since certification, dives in last year, number of diving days, number of dives in a repetitive series, last dive depth, nitrox use, and drysuit use.
Bestneighbor: Risk factors include: high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, and excessive alcohol, among others.
S4: The momentum of the water is suddenly transferred into the fitting and Newton's Third Law kicks in forming growing the high-pressure region of water as it all "piles up" in the pipe.
Bestneighbor: When properly confined and close to the surface it can periodically release some of the built-up pressure in eruptions of hot water and steam that can reach up to 390 feet.
S5: These included the launch of a national campaign to promote interfaith harmony, the proposal of legislation to ban hate speech and related literature, the proposed introduction of comparative religion as a curriculum subject, the introduction of quotas for religious minorities in government posts and the reservation of four Senate seats for minorities.
Bestneighbor: The Constitution provides for freedom of religion and does not establish a state religion; however, in practice the Government imposes legal restrictions on all forms of religious expression.
S1: Nevertheless, the complete characterization of the stable throughput region for such system is notoriously difficult, since the computation of the steady state distribution of the two-dimensional Markov chain (mc) model for both finite queues is prohibitively complex.
Bestneighbor: To this end, we propose a hierarchical Dirichlet process hidden Markov model (hdp-hmm), combined with wide-sense stationary time series spectral estimation to construct a generative model for personalized subject sleep states.
S2: Filter output truncation can reduce the overhead by discarding the filter tails but may also significantly destroy the orthogonality of fbmc system, by introducing inter carrier interference and inter symbol interference terms in the received signal.
Bestneighbor: Next, by taking carrier frequency offset, timing offset, insufficient guard interval between symbols and filter tail cutting into consideration, an analytical system model is established.
S3: using a sample of galaxies selected from the sloan digital sky survey data release 7 (sdss dr7) and a catalog of bulge-disk decompositions, we study how the size distribution of galaxies depends on the intrinsic properties of galaxies.
Bestneighbor: Here we use cosmological models of galaxy formation to show that m31's massive and metal-rich stellar halo, containing intermediate-age stars, dramatically narrows the range of allowed interactions, requiring a single dominant merger with a large galaxy about 2 gyr ago.
S4: In this paper, we bring these two separate strands of research together to systematically explore links between internal migration and education across a global sample of 57 countries at various stages of development, using data drawn from the ipums database.
Bestneighbor:The main methodological principles of the environment design and development are considered among them there are the principles of open education, open science and also the specific principles inherent rightarrow the cloud-based systems.
S5: This paper is devoted to the asymptotic analysis of a thin film equation which describes the evolution of a thin liquid droplet on a solid support driven by capillary forces.
Bestneighbor: Using landau-ginzburg-devonshire theory, we considered the impact of the flexoelectro-chemical coupling on the size effects inpolar properties and phase transitions of thin ferroelectric films with a layer of elastic defects.

We also show the model interpretability on sentence representation by visualizing the topic proportions, where the top five topics of a sentence are represented by ranked words in the dictionary. In additional to this, for each sentence, we display the six nearest words, which are ranked by the cosine distances computed with the corresponding topic distributions. We select four sentences from MedicalTS, and run the bi-DCSR method to obtain the representation of each sentence. Then, the interpretable topic proportions with *beta* and the nearest words with ϕ are output. Table 8 shows the results. The four sentences belong to Cardiovascular/Pulmonary, Orthopedic, Surgery, and Neurology, respectively.

From Table 8, we find that the topics are clustered around a certain hidden semantic, e.g., the topics describing "surgery" in row 3. This phenomenon can be found in Fig. 8d. By embedding the sentences and the words in the same topic space, we can easily show the representation interpretability on hidden semantic spaces. This is an important benefit of choosing topic modeling to learn sentence representations, and it is crucial for certain real applications on some closed domains, e.g., intention detection on medical dialog and Question–Answer systems.



(a) Wikipedia (b) arXiv

Fig. 11. The plots of the selected sentences and the nearest words. Each plots denotes a word, and a star denotes a sentence described in Table 7.

Table 8
Visualization of the topic proportions of four sentences from MedicalTS with the top five topics, ranked by the probabilities of the topics. Sentence labels and the six nearest words are provided.

Sentences	Top five topic proportions	The six nearest words
The patient underwent an echocardiogram, which shows severe mitral regurgitation and also pleural effusion.	0.76429 ['cardiovascular', 'pulmonary', 'patient', 'chest', 'disease', 'daily'] 0.15675 ['history', 'patient', 'past', 'normal', 'blood', 'family'] 0.04422 ['pain', 'negative', 'denies', 'history', 'breath', 'shortness'] 0.03432 ['left', 'coronary', 'artery', 'valve', 'aortic', 'vessel'] 0.00009 ['disc', 'central', 'normal', 'spine', 'canal', 'cervical']	cardiovascular, breath, diverticulitis, patient, urgency, pain
He came down on another player's foot sustaining what he describes as an inversion injury.	0.70033 ['orthopedic', 'pain', 'joint', 'left', 'back', 'knee'] 0.13385 ['patient', 'pain', 'history', 'time', 'states', 'past'] 0.07312 ['foot', 'metatarsal', 'patient', 'incision', 'proximal', 'left'] 0.04664 ['tendon', 'shoulder', 'tear', 'cuff', 'rotator', 'normal'] 0.02128 ['fracture', 'wound', 'left', 'reduction', 'distal', 'open']	orthopedic, patient, pain, knee, bone, shoulder
A Freer was then inserted beneath the ligament, and dissection was carried out proximally and distally.	0.30681 ['surgery', 'left', 'patient', 'diagnoses', 'noted', 'procedure'] 0.28719 ['carpal', 'tunnel', 'nerve', 'ulnar', 'transverse', 'ligament'] 0.25254 ['surgery', 'patient', 'procedure', 'anesthesia', 'diagnosis', 'room'] 0.07869 ['surgery', 'knee', 'femoral', 'removed', 'patient', 'position'] 0.07314 ['surgery', 'skin', 'incision', 'suture', 'patient', 'left']	surgery, procedure, carpal, patient, tunnel, knee
Today, the patient returns rightarrow clinic because of acute onset of headaches that she has had since her shunt was adjusted after an MRI on 04/18/08.	0.50137 ['neurology', 'patient', 'history', 'time', 'day', 'years'] 0.33964 ['neurology', 'left', 'bilaterally', 'normal', 'noted', 'patient'] 0.15804 ['patient', 'pain', 'history', 'time', 'states', 'past'] 0.00002 ['disc', 'central', 'normal', 'spine', 'canal', 'cervical'] 0.00002 ['fracture', 'wound', 'left', 'reduction', 'distal', 'open']	neurology, patient, headaches, pain, history, spine

8. Conclusion

In this paper, we investigate the issue of the sentence representation learning on closed domains, where the data is deficient in training a deep learning model from scratch, and the sentence representation interpretability is also required. To handle this problem, we introduce a novel and unsupervised Bayesian framework to train and infer sentence representations on closed domains. The method can generate high-quality sentence representations by using rich contextual elements in a piece of text. In addition, the HPTM is proposed to support the framework, where we embed the words and sentences in the same topic space. The proposed method aims to tackle the issue of sentence representation learning, especially for closed domains with fewer data. The strength of it includes that the proposed method can efficiently learn high-quality sentence representations in such closed domains from scratch, and it is also easy to train and inference. This proposed method is flexible and expandable for topic models. Compared with the state-of-the-art unsupervised learning approaches, our method achieves better performance. Also, built on topic modeling, the proposed model is generalizable and easily interpretable in contrast to deep learning models. Future work is going to focus on reducing the complexity of the proposed model for large-scale corpora, making it faster to train and inference in open domains.

CRediT authorship contribution statement

Shuangyin Li: Conceptualization, Methodology, Software, Writing - original draft. **Weiwei Chen:** Software, Visualization, Data curation, Resources. **Yu Zhang:** Writing - review & editing, Validation. **Gansen Zhao:** Supervision, Validation, Writing - review & editing, Funding acquisition. **Rong Pan:** Supervision. **Zhenhua Huang:** Writing - review & editing, Validation. **Yong Tang:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62006083), GuangZhou Basic and Applied Basic Research Foundation (202102020654) and Applied Basic Research Fund of Guangdong Province (2019B1515120085). Gansen Zhao and Zhenhua Huang is the co-corresponding authors.

References

[1] M. Tan, C. Dos Santos, B. Xiang, B. Zhou, Improved representation learning for question answer matching, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 464–473.
[2] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.
[3] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
[4] R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: Advances in neural information processing systems, 2015, pp. 3294–3302.
[5] L. Logeswaran, H. Lee, An efficient framework for learning sentence representations, arXiv preprint arXiv:1803.02893 (2018).
[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, arXiv preprint arXiv:1906.08237 (2019).
[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[9] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, M. Baroni, What you can cram into a single vector: Probing sentence embeddings for linguistic properties, ACL (2018).
[10] J. Chi, J. Ouyang, C. Li, X. Dong, X. Li, X. Wang, Topic representation: Finding more representative words in topic models, Pattern Recognition Letters 123 (2019) 53–60.
[11] S. Li, R. Pan, H. Luo, X. Liu, G. Zhao, Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling, Knowledge-Based Systems 106827 (2021).
[12] S. Li, Y. Zhang, R. Pan, K. Mo, Adaptive probabilistic word embedding, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 651–661.
[13] S. Li, Y. Zhang, R. Pan, M. Mao, Y. Yang, Recurrent attentional topic model, in: 31th AAAI Conference on Artificial Intelligence (AAAI-17), in: Association for the Advancement of Artificial Intelligence(AAAI), 2017, pp. 3223–3229.
[14] M. Selvi, K. Thangaramya, M. Saranya, K. Kulothungan, S. Ganapathy, A. Kannan, Classification of medical dataset along with topic modeling using ldas, in: Nanoelectronics, Circuits and Communication Systems, Springer, 2019, pp. 1–11.
[15] J. Tang, Z. Meng, X. Nguyen, Q. Mei, M. Zhang, Understanding the limiting factors of topic modeling via posterior contraction analysis, in: International Conference on Machine Learning, 2014, pp. 190–198.
[16] J. Chen, Z. Gong, W. Liu, A nonparametric model for online topic discovery with word embeddings, Information Sciences 504 (2019) 32–47.
[17] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda, G.L. Pappa, A general framework to expand short text for topic modeling, Information Sciences 393 (2017) 66–81.
[18] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi, J. Ouyang, Filtering out the noise in short text topic modeling, Information Sciences 456 (2018) 83–96.
[19] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365 (2018).
[20] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation., in: EMNLP, Vol. 14, 2014, pp. 1532–1543.
[21] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv:1301.3781.
[22] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R.S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).
[23] J. Wieting, K. Gimpel, Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations, ACL (2018).
[24] W. Xu, S. Li, Y. Lu, Usr-mtl: an unsupervised sentence representation learning framework with multi-task learning, Applied Intelligence (2020), https://doi.org/10.1007/s10489-020-02042-2.
[25] M. Pagliardini, P. Gupta, M. Jaggi, Unsupervised learning of sentence embeddings using compositional n-gram features, NAACL (2018).
[26] R. Socher, E.H. Huang, J. Pennin, C.D. Manning, A.Y. Ng, Dynamic pooling and unfolding recursive autoencoders for paraphrase detection, in: Advances in neural information processing systems, 2011, pp. 801–809.
[27] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, 2018.
[28] T. Zhao, K. Lee, M. Eskenazi, Unsupervised discrete sentence representation learning for interpretable neural dialog generation, ACL (2018).
[29] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[30] C. May, A. Wang, S. Bordia, S.R. Bowman, R. Rudinger, On measuring social biases in sentence encoders, arXiv preprint arXiv:1903.10561 (2019).
[31] F. Tian, B. Gao, D. He, T.-Y. Liu, Sentence level recurrent topic model: letting topics speak for themselves, arXiv preprint arXiv:1604.02038 (2016).
[32] S. Li, Y. Zhang, R. Pan, Bi-directional recurrent attentional topic model, ACM Trans, Knowl. Discov. Data 14 (6) (2020).
[33] T. Iwata, T. Hirao, N. Ueda, Topic models for unsupervised cluster matching, IEEE Transactions on Knowledge and Data Engineering 30 (4) (2017) 786–795.

[34] S. Li, H. Luo, G. Zhao, bi-hptm: An effective semantic matchmaking model for web service discovery, in: 2020 IEEE International Conference on Web Services (ICWS), 2020, pp. 433–440.
[35] S. Li, H. Luo, G. Zhao, M. Tang, X. Liu, bi-directional bayesian probabilistic model based hybrid grained semantic matchmaking for web service discovery, World Wide Web 25 (2) (2022) 445–470.
[36] P. Gupta, Y. Chaudhary, F. Buettner, H. Schütze, Document informed neural autoregressive topic models with distributional prior, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6505–6512.
[37] J.-T. Chien, C.-H. Lee, Deep unfolding for topic models, IEEE transactions on pattern analysis and machine intelligence 40 (2) (2017) 318–331.
[38] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.
[39] S. Patterson, Y.W. Teh, Stochastic gradient riemannian langevin dynamics on the probability simplex, in: Advances in neural information processing systems, 2013, pp. 3102–3110.
[40] O. Bousquet, L. Bottou, The tradeoffs of large scale learning, in: NIPS, 2008.
[41] P. Liang, D. Klein, Online em for unsupervised models, ACL (2009) 611–619.
[42] M. Hoffman, F.R. Bach, D.M. Blei, Online learning for latent dirichlet allocation, in: advances in neural information processing systems, 2010, pp. 856–864.
[43] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring topic coherence over many models and many topics, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 2012, pp. 952–961.
[44] O. Levy, Y. Goldberg, I. Dagan, Improving distributional similarity with lessons learned from word embeddings, Transactions of the Association for, Computational Linguistics 3 (2015) 211–225.
[45] M. Yang, T. Cui, W. Tu, Ordering-sensitive and semantic-aware topic modeling, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[46] L. Du, W. Buntine, H. Jin, C. Chen, Sequential latent dirichlet allocation, Knowledge and information systems 31 (3) (2012) 475–503.
[47] H. Larochelle, S. Lauly, A neural autoregressive topic model, Advances in Neural Information Processing Systems (2012) 2708–2716.
[48] H. Zhao, L. Du, W. Buntine, G. Liu, Metalda: A topic model that efficiently incorporates meta information, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 635–644.
[49] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, Y. Goldberg, Fine-grained analysis of sentence embeddings using auxiliary prediction tasks, ICLR (2017).