BCTM: A Topic Modeling Method Based on External Information

Gang Liu[1,3], Taiying Wan[1,3](✉), Jinfeng Yu[1,3], Kai Zhan[2], and Wei Wang[1,3]

[1] College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
wantaiying@hrbeu.edu.cn
[2] PwC Enterprise Digital, PricewaterhouseCoopers, Sydney, NSW 2070, Australia
[3] National Engineering Laboratory of E-Government Modeling Simulation, Harbin Engineering University, Harbin 150001, China

Abstract. Topic models are often used as intermediate algorithms for text mining and semantic analysis in natural language processing, and have a wide range of functions. However, most of the existing improvements to the topic model use word embedding to improve the accuracy of text modeling, but ignore the external information in the text. This paper proposes a topic model BCTM (Bi-Concept Topic Model) using the word feature information and concept information. Based on the BTM topic model, BCTM introduces word feature information through word vector technology and concept information based on ConceptNet to optimize topic modeling. The construction method of Bi-Concept pair is proposed. Based on ConceptNet semantic network, and the content of text is enriched with concept information. A more accurate topic distribution is obtained through the improved topic model, at the same time, due to the rich feature information, the model is also superior to the baseline model in short text modeling. The experiments prove that the bilingual topic model proposed in this paper has a good performance in modeling accuracy.

Keywords: Topic modeling · Word embedding · External information

1 Introduction

The study of natural language is no longer limited to linguistics, history and other liberal arts fields, the use of computers to deal with natural language is an important development direction of artificial intelligence [1]. With the rapid development of the Internet, a large number of text data are generated in social networks, online shopping, news sites and other information flow sites, which are difficult to find in traditional information sources, so there is a demand for unsupervised topic analysis. Probability-based topic models such as LDA [2] are common methods for this task. Topic models are widely used and are often used to extract various text features [3].

W. Wang and J. Wu (Eds.): BROADNETS 2023, LNICST 511, pp. 95–110, 2023.
https://doi.org/10.1007/978-3-031-40467-2_6

In recent years, improving the topic model and making it suitable for short text mining has become a hot research direction in the field. Some researchers improve the modeling quality of short text by expanding text information, such as aggregating short text into pseudo-documents and extracting features by topic modeling based on pseudo-document expectations. Some scholars assume that the number of topics in the short text is sparse and the document is constrained in a small number of topics, but this method is not suitable for situations where the short text may cover multiple topics. But intuitively, because the aggregation of pseudo-documents will inevitably bring noise, it is difficult to ensure the quality of topics based on pseudo-documents, so it is necessary to take other ways to deal with "noise reduction".

External information is a good way to enrich text information. The external information in this paper is embodied in conceptual information and word feature information. Word vector techniques such as Glove [4] can bring great help to topic semantic enhancement, because the pre-trained word vector model can well supplement the sparse features of short texts. In addition to the frequently used word vectors, external information such as author tags and timestamps of the text can theoretically be used to improve the quality of short text modeling.

In summary, this paper proposes an optimized BTM improved model BCTM,. The main contributions are as follows:

The model overcomes the shortcomings of the previously mentioned model, such as complex structure, non-conjugation and single channel of information acquisition, by transforming meta-information into word label information. The tag information is independent of the model itself so that words with similar tags have similar distribution weights on the topic.

This paper puts forward the construction method of Bi-Concept pair, which enriches the content of the text by introducing conceptual information, so as to construct an effective topic model on the short text.

Several groups of experiments were carried out with confusion degree and theme consistency as evaluation indexes. Experiments show that compared with LDA and other models, the BCTM model based on external information performs better under the same conditions.

2 Related Work

2.1 Topic Model

Since David Blei and others put forward the classical probabilistic topic model LDA, scholars in the field have strong interest and enthusiasm on how to improve this model. LDA belongs to a three-layer Bayesian network structure and has good scalability. Because LDA belongs to the word bag model, it does not consider the order of words and context, which provides a lot of room for the improvement of LDA [5].

The traditional LDA topic model does not perform well in short texts, but its defect is that it does not have enough lexical information to make its statistical meaning valid. How to enrich document information and apply it to short text topic modeling is one of the hot research directions of scholars in recent years.

Document aggregation is an idea of information expansion based on the original text. Zhao [6] proposed a generation model, which aggregates short texts into clusters by using relevant meta-information, which alleviates the poor modeling effect of LDA in short texts. From the point of view of words, Yan [7] proposed a topic model BTM (Biterm Topic Model), which constructs short texts as two-word (biterm) sets. The model makes up for the text sparsity of short texts, and its sampling is based on biterm pairs, which enhances the word co-occurrence of short texts at the document level. In recent years, researchers have made an endless stream of studies on the improvement of classical models. For example, Wu [8], Zhu [9], Li [10] and Huang [11] have made improvements based on the BTM model, but the short text aggregates long text at the same time can not avoid noise, the accuracy of this method is difficult to be effectively guaranteed.

The general idea of introducing word vector feature is to increase the accuracy of topic-word distribution by making words with similar semantics more likely to be assigned to the same topic. Nguyen [12] proposed a LF-LDA (Latent-Feature LDA) model, which combines the word vectors trained by the external corpus into the topic modeling, and extends the topic model to change the Dirichlet component at the topic-word level into a mixture of the word vector and the original subject word distribution. Li [13] proposed the GPU-PDMM model. In the sampling process, the generalized P ó lya urn (GPU) model is used to promote the semantic related words under the same topic, and through the GPU model, the background knowledge about the semantic relations of words learned from millions of external documents can be easily used to improve the topic modeling of short text. Gao [14] proposed the CRFTM (Conditional Random Field regularized Topic Model) model, which uses a conditional random field regularization model so that related words can share the same topic assignment. At the same time, such methods only rely on word vectors to enrich word co-occurrence information, while other external information such as conceptual information between words are not effectively used, so although the accuracy of the model is higher than that of LDA model, there is still room for improvement.

2.2 External Information

The existing improved topic models often ignore the rich external information in the corpus, which also provides a new direction for the research of topic modeling in this paper. External information refers to the information outside the content of the text. Compared with the independent text, the word feature information of the text vocabulary, the associated knowledge information, the label of the text label and so on belong to the external information. The introduction of external information increases the access to information, which can effectively improve the accuracy of modeling.

3 BCTM Model Based on External Information

3.1 Overview of the Model

On the basis of the BTM model, the BCTM model proposed in this paper mainly constructs the short text topic model from the following points, and the specific probability model is shown in Fig. 1.

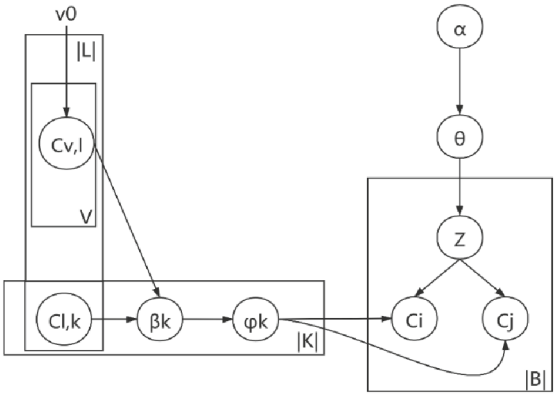

Fig. 1. BCTM probability model diagram

The main contents are as follows:

(1) the word feature information is introduced by the way of label. Lexical features are introduced as external information as a priori in order to make words with similar features have similar Dirichlet prior parameters under the same topic, which means that their probability of appearing in a certain topic is similar. Lexical tags are obtained by binarization of word vectors. At the same time, they can be used as word feature tags according to the way they are generated and combined with conceptual features.
(2) introducing conceptual information. The conceptual information is introduced into the conceptual knowledge network ConceptNet, and the algorithm is designed to improve the biterm generation process of BTM. The Bi-Concept pairs are constructed and sampled on the Bi-Concept set.

As shown in Fig. 1, it is assumed that the corpus set D consists of D documents, in which the vocabulary of each document $d \in \{1, \ldots, D\}$, corpus consists of one word. The vocabulary of each document is processed with the help of the conceptual network to get the Bi-Concept pair set B, and the constructed Bi-Concept pair is the $\{C_i, C_j\}$, concept $n \in \{C_1, \ldots C_i, C_j, \ldots C_n\}$. The arrow in the figure shows the influence relationship on the subsequent parameters. For example, the Vocabulary-word feature label matrix of the word $C_{v,l}$ is deduced according to the distribution of the word v0, and thus affects the Dirichlet priori hyperparameter β_k. BCTM models the topic on the constructed Bi-Concept set, including K topics. V_0, α and β_k are the prior parameters, and the word feature label is L. the generation process is as follows:

(1) for each topic K:
  a. For word feature tags $l \in |L|$, sampling gamma distribution $C_{v,l} \sim Ga(v_0, v_0)$
  b. For the word v, calculate the $\beta_{k,v} = \prod_{l=1}^{L_{word}} c_{l,k}^{C_{v,l}}$
  c. Sampled Dirichlet distribution $\varphi_k \sim Dir(\beta_k)$
(2) for concept pair sets:
  a. Sampled Dirichlet distribution $\theta \sim Dir(\alpha)$
  b. For topic Z, sample category distribution $Z \sim Cat(\theta)$
  c. For conceptual C_n, sampling category distribution $C_n \sim Cat(\varphi_z)$

The symbols and explanations related to the structure of the model are shown in Table 1. It is worth mentioning that in order to integrate the word feature information into the model, the BCTM model uses binary label information to learn a specific Dirichlet prior β_k. That is to say, for any topic extracted from the BCTM model, if the characteristics of the C_i, C_j are similar, then the numerical representation of the corresponding Dirichlet prior β_k on the topic should also be close. This can be understood as: the probability of selecting words with similar features on the same topic is about the same, which is the function of introducing word feature tags. In the end, the sampling of the BCTM model is carried out in the whole set of concept pairs. By constituting the short text into a concept pair set, not only the sparse word co-occurrence information of the original short text can be enriched, but also the accuracy of model modeling can be improved after the introduction of conceptual network.

Table 1. Symbol and meaning of BCTM model

Symbol	Meaning
B	Concept pair set
V	Glossary
K	Number of topics
L	Label dimension
Z	Concept to collection topic
α	Dirichlet prior parameter
θ	The topic distribution of concepts to collections
C_i, C_j	Concept pair
φ_k	Word distribution of topic k
β_k	Dirichlet prior parameter
$C_{l,k}$	Topic-word feature label relevance weight
$C_{v,l}$	Vocabulary-word feature label matrix
v_0	Super parameter

3.2 Introduction of Word Feature Information

At present, the word vectors used in most models can effectively measure the similarity or potential distance between words, but the word vectors can not give the representation of the relationship between words. In order to solve this problem, this paper proposes a binarization word feature label method, which can use word features as external information tags, so that the model sampling process can affect the specific Dirichlet prior corresponding to a topic through labels. In this way, the distribution of words under a topic has a certain law, that is, words with a priori approximation are more likely to appear in the same topic distribution. And then improve the theme consistency of the topic model.

Algorithm 1 word feature label generation algorithm

Input: The word vector set corresponding to the pre-trained vocabulary$G = \{V_1, V_2, \cdots, V_{Count}\}$, where the word vector $V_{count} = \{r_1, r_2, \cdots, r_n\}$
Output:Binary word feature tag set $C' = \{C'_1, C'_2, \ldots, C'_{Count}\}$, wherein, word feature tag $C'_i = \{c_1, c_2, \ldots, c_n\}$
For $i = 1,2, \ldots, Count$ Do
  For $j = 1,2, \ldots, n$ Do
    IF $r_j > 0$ Do
   Word vector positive summation$S_+ += r_j$
    ELSE IF $r_j < 0$ Do
   Summation of negative values of word vectors$S_ += r_j$
    End IF
  End For
   Calculate the positive average $M_+ = S_+/n$ of the word vector V_i
   Calculate the negative average $M\_ = S\_/n$ of the word vector V_i
  For $j = 1,2, \ldots, n$ Do
    IF $r_j > M_+$ Do
   Current dimension word feature label value$c_j = 1$
      ELSE IF $r_n < M\_$ Do
   Current dimension word feature label value$c_j = 1$
    ELSE Do
   Current dimension word feature label value$c_j = 0$
    End IF
  End For
End For

First, the average M_+, M_-, of the word vector is calculated, where M_+ is the average value of the sum of all positive elements in the word vector, and M_- is the average value of the sum of all negative elements in the word vector. Then, according to whether the value of each dimension of the word vector is positive or negative, it is compared with the calculated average M_+, M_-. According to the comparison of the value of each

dimension of the word vector with the calculated average, the operation is performed in turn:

(1) if the current dimension value $> M_+$, the current dimension value $= 1$.
(2) if the current dimension value is less than M_-, the current dimension value is 1.
(3) if it is otherwise, the current dimension value is 0.

The main idea of this method is to retain the "prominent" part of the word vector feature as much as possible, while the more prominent feature refers to the corresponding dimension of the word vector whose value is greater than the average. In this way, the more important features can be retained as much as possible, and the weaker features can be removed at the same time, which can screen the word features to a certain extent and retain the word feature information with good quality. Suppose that the size of the text vocabulary is Count, and the n-dimensional word vector is expressed in the form of $V_i = \{r_1, r_2, \ldots, r_n\}$. In order to make the word feature information can be applied to the model, the binary word feature tag generation algorithm proposed in this paper is shown in Algorithm 1.

At present, many short text topic models combined with word vectors are modeled on the word vector space. Different from those short text topic models, the BCTM model uses word vector tools and introduces word feature information in the form of tags on the basis of the classical topic model to make up for the sparse word co-occurrence of short texts.

3.3 Introduction of Conceptual Information Based on ConceptNet

BCTM model improves the accuracy of topic modeling by introducing concepts as external information. The purpose of constructing Bi-Concept set based on ConceptNet for topic modeling is to improve the quality of the aggregated corpus through high-quality a priori knowledge. However, for the construction of the Bi-Concept collection, we should also pay attention to how to build the concept pair, and the extent to which the concept pair of construction should be extended should also be taken into account. The following will give the algorithm description of the construction of the Bi-Concept used in this paper.

4 Derivation of Parameters

Because the BCTM model introduces the word feature information to sample a corresponding Dirichlet prior for each word. Different from the BTM model, which can not generate document-topic distribution due to the sampling of word pairs, BCTM model modeling is based on the collection of concept pairs, which is equivalent to sampling on the mixed text composed of concept pairs, so it is similar to the sampling process of LDA model.

Given the topic distribution $\theta_{1:D}$ corresponding to a document and the joint distribution of the corresponding word distribution $\varphi_{1:K}$, BCTM model under the topic, the joint distribution can be shown by formulas 1 and 2.

$$\mathbf{P}\left(\mathbf{w_{1:D}}, \mathbf{z_{1:D}}|\boldsymbol{\theta}_{\mathbf{1:D}}, \boldsymbol{\varphi}_{\mathbf{1:K}}\right) = \prod_{\mathbf{d}=1}^{\mathbf{D}} \prod_{\mathbf{k}=\mathbf{1}}^{\mathbf{K}} \boldsymbol{\theta}_{\mathbf{d},\mathbf{z_{d,i}}} \boldsymbol{\varphi}_{\mathbf{z_{d,i}},\mathbf{v}} \tag{1}$$

$$\prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,z_{d,i}} \varphi_{z_{d,i},v} = \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{m_{d,k}} \prod_{d=1}^{D} \prod_{k=1}^{K} \varphi_{k,v}^{n_{k,v}} \tag{2}$$

As shown in the above formula, the modeling method for the corpus set D d $\in$ {1, .., D} model is to generate the topics of the document BCTM from the number of topics K, where each topic k $\in$ {1, . . . , K} chooses to generate words from the vocabulary V according to the specific distribution. The BCTM model first samples the topic $z_{d,i} \in \{1, \ldots, K\}$ according to the document-topic distribution $\theta_d \in \mathbb{R}_+^K$, and then samples the lexical $w_{d,i}$ according to the topic-word distribution $\varphi_{z_{d,i}}$. Here $n_{k,v}$ represents the number of v words assigned to topic k, while $m_{d,k}$ refers to the total number of words assigned to topic k by document d. The LDA model uses Dirichlet conjugation to solve the probability distribution, and the solution method here is similar, as shown in formula 3.

$$P(w_{1:D}, z_{1:D}|\alpha_{1:D}, \beta_{1:K}) = \prod_{d=1}^{D} \frac{\boldsymbol{Beta_K}(\boldsymbol{\alpha_d} + \boldsymbol{m_d})}{\boldsymbol{Beta_K}(\boldsymbol{\alpha_d})} \prod_{k=1}^{K} \frac{\boldsymbol{Beta_V}\left(\boldsymbol{\beta_k} + \boldsymbol{n_k}\right)}{\boldsymbol{Beta_V}\left(\boldsymbol{\beta_k}\right)} \tag{3}$$

Algorithm 2 Bi-Concept pair construction algorithm

Input: The D= $\{d_1, d_2, \ldots, d_D\}$, threshold N of the document set to be constructed, and the word collection $V = \{v_1, v_2, \ldots, v_n\}$ of the concept network node set $Node = \{node_1, node_2, \ldots, node_{num}\}$, document d_D

Output: Bi-Concept collection $B = \{b_1, b_2, \ldots, b_D\}$ built by document set D

For $i = 1,2, \ldots, D$ **Do**
  For $j = 1,2, \ldots, n-1$ **Do**
    Find the starting node $node1$ of v_j corresponding to the conceptual network
    Traversing the conceptual network, constructing the node set $Node' = \{node'_1, node'_2, \ldots, node'_{num}\}$ of $node1$ which can be reachable according to the edge step in the conceptual network.
    For $k = j+1, j+2, \ldots, n$ **Do**
      Find the starting node $node2$ of v_k corresponding to the conceptual network
      IF Both $node1$ and $node2$ are in node set b_i **Do**
        Form a Bi-Concept pair of $node1$ and $node2$, and put them into set b_i.
        Put v_j and v_k into set V_C
      End IF
    End For
  End For
  IF The Bi-Concept pair in set b_i is less than the threshold N **Do**
    Remove the vocabulary that appears in V_C in V, and get the vocabulary set $V' = \{v'_1, v'_2, \ldots, v'_m\}$ that does not participate in the formation of Bi-Concept pairs
    For $l = 1,2, \ldots, m-1$ **Do**
      Find the starting node $node1'$ of the conceptual network corresponding to v_l
      Traverse the conceptual network and construct the node set $Node'' = \{node''_1, node''_2, \ldots, node''_{count}\}$ that $node1'$ can reach within two steps according to the edge in the conceptual network
      For $g = l+1, l+2, \ldots, m$ **Do**
        Find the starting node $node2'$ of the conceptual network corresponding to v_g
        IF Both $node1'$ and $node2'$ are in node set $Node''$ **Do**
          Combine $node1'$ and $node2'$ into a Bi-Concept pair and put them into set b_i
        End IF
      End For
    End For
  End IF
End For

As shown in formula 3, $\Gamma(\cdot)$ is the gamma function, and $\mathrm{Beta}_N(\cdot)$ is the N-dimensional beta function shown in formula 4.

$$\boldsymbol{Beta}_N(\boldsymbol{x}) = \frac{\prod_n \boldsymbol{\Gamma}(\boldsymbol{x}_n)}{\boldsymbol{\Gamma}\left(\sum_n \boldsymbol{x}_n\right)} \tag{4}$$

Assuming that the Dirichlet prior and β for different documents and topics are known, the probability calculation for the assigned topic $z_{d,i}$ d is shown in Formula 5.

$$\boldsymbol{P}(\boldsymbol{z}_{\boldsymbol{d},i} = \boldsymbol{k} | \boldsymbol{z}_{\boldsymbol{1:d}}^{-\boldsymbol{z}_{d,i}}, \boldsymbol{w}_{\boldsymbol{1:D}}, \boldsymbol{\alpha}_{\boldsymbol{1:D}}, \boldsymbol{\beta}_{1:\boldsymbol{K}}) \propto \left(\boldsymbol{\alpha}_{\boldsymbol{d},\boldsymbol{k}} + \boldsymbol{m}_{\boldsymbol{d},\boldsymbol{k}}\right) \frac{\boldsymbol{\beta}_{\boldsymbol{k},\boldsymbol{v}} + \boldsymbol{n}_{\boldsymbol{k},\boldsymbol{v}}}{\boldsymbol{\beta}_{\boldsymbol{k},\cdot} + \boldsymbol{n}_{\boldsymbol{k},\cdot}} \tag{5}$$

5 Experimental Results and Analysis

As the BCTM model is conceived and implemented with short texts as the main corpus, the corpus adopted in the experiment are all short texts on social media, such as online text fragments, blog content and so on. So far, there is no open and authoritative explanation for the clear definition of short texts, so the experimental corpus is chosen as shorter texts with an average of 15 words, short texts with an average of 100 words, and general texts with an average of more than 1000 words. The reason why we use the dataset with long average vocabulary is that the modeling quality of BCTM model can be verified only from the point of view of short text, although it can reflect the effect of the model to some extent, but because BCTM introduces word feature information and concept information, BCTM model can also achieve good modeling quality in general text modeling. Only from the perspective of short text, there may be some limitations in the quality evaluation of the model, so in the experimental part of the text, several public data sets of three lexical intervals are selected for topic modeling experiments. The specific dataset is described below.

WS (Web-Snippet) [15] dataset. The dataset is widely used in short text topic modeling testing. Contains 12237 Web search fragments. The vocabulary of the dataset contains 10052 words, with an average of 15 words in each segment.

KOS [16] dataset. The data set is obtained from the UCI machine learning database and mainly contains some blog entries and so on. Archambeau et al. used this data set in the experimental part of the model in which the implicit IBP Dirichlet distribution was introduced. It has 3430 documents with a vocabulary of 610 and a vocabulary of 677, with an average of about 100 words per text.

NIPS dataset. The data is also downloaded from UCI machine learning database, and its content is related to NIPS papers. The dataset has 1500 texts, and the glossary contains 12419 words, with an average of about 1266 words in each text.

StackOverFlow [17] dataset. The data set contains 20000 question-and-answer sentences on the StackOverFlow technical question-and-answer website, which are divided into 20 categories for text classification experiments, with an average of about 100 words per text.

Sogou news data set. The data set is the news corpus of Sogou Lab. A total of 1000 texts are selected from each of the five categories of finance and economics, health,

Table 2. Test data set

Data set	Document number	Average document length(words)
Web-Snippet	12237	15
KOS	3430	100
NIPS	1500	1266
StackOverFlow	20000	100
Sogou	1000	80

education, military and culture as a set of training. It is mainly used in text classification experiment, with an average of about 80 words per text (Table 2).

Contrast experiment setting of topic modeling.

The BCTM proposed in this paper is compared with the following methods on the above different data sets.

LDA model, LDA as the most classic topic model, so far there have been many scholars on this basis to improve the model. However, it can not be ignored that the experiment of LDA model on short text and general text is very necessary, because the experimental results are often used as a benchmark to compare with other models. Therefore, this paper also uses LDA as one of the baseline models.

BTM topic model. The most prominent aspect of the BTM model is that it puts forward the idea of transforming the short text corpus into biterm. Different from LDA model, BTM model is sampled based on biterm, so its document-topic distribution needs to be extrapolated to biterm. The advantage of BTM is that it expands the sparse content of short text by constructing biterm collection. Theoretically, the performance of short text should be better than that of LDA model. This paper also adopts the idea of word pairs, so BTM model is also one of the baseline models for comparison.

LF-LDA topic model. The LF-LDA model also introduces the word vector as a supplement to the model to improve the LDA topic model. In the process of topic modeling, LF-LDA introduces the word vector information trained in the large corpus to make the topic randomly select the characteristics of the original text or word vector with a specific distribution, so as to improve the topic consistency of the topic model. This paper also introduces a lot of external information to influence the Dirichlet priori of the text, so the LF-LDA model is also one of the baseline models for comparison.

In this experiment, two mainstream indicators are used to evaluate the quality of a topic model, that is, confusion evaluation [18], topic consistency [19]. Next, we will introduce these evaluation criteria in detail:

The degree of confusion is used to measure the quality of the sample predicted by the probability model. It is an important evaluation index in the field of topic model. The smaller the value of confusion is, the higher the accuracy of the model is.

The confusion evaluation experiments are carried out on three data sets WS, KOS and NIPS, and the BCTM model is compared with the above three models. The results of the comparison are shown in Table 3 below.

Table 3. Comparison of Perplexity of each Model

Data set/$K = 50$	LDA	BTM	LF-LDA	BCTM
WS	923.57	417.98	763.29	**391.61**
KOS	564.58	217.92	286.29	**206.94**
NIPS	431.41	273.23	1098.24	280.34
Data set/$K = 100$	LDA	BTM	LF-LDA	BCTM

(*continued*)

Table 3. (*continued*)

Data set/$K = 50$	LDA	BTM	LF-LDA	BCTM
WS	629.54	302.99	655.81	312.89
KOS	512.38	180.86	269.61	**162.38**
NIPS	431.46	265.71	1083.42	267.53
Data set/$K = 150$	LDA	BTM	LF-LDA	BCTM
WS	501.98	292.41	570.76	**255.24**
KOS	502.54	165.61	231.56	**147.09**
NIPS	431.72	265.70	1074.02	**262.17**
Data set/$K = 200$	LDA	BTM	LF-LFA	BCTM
WS	436.03	236.15	512.81	**198.18**
KOS	496.38	167.29	229.54	**144.01**
NIPS	440.23	265.70	1007.51	**259.70**

From the above table, we can see that on the three data sets of WS, KOS and NIPS, the number of different topics and the degree of confusion calculated by different models are also different. In the table, the models with the least confusion under the same dataset and the same number of topics have been boldly marked. On the whole, the confusion label of BCTM model is relatively low, and the confusion performance of modeling with 50 topics on NIPS data sets, 100 topics on WS data sets and 100 topics on NIPS data sets is not the best. Correspondingly, the best text in these cases is the BTM model. The visualization of the confusion degree experiment on the three corpus is shown in Fig. 2.

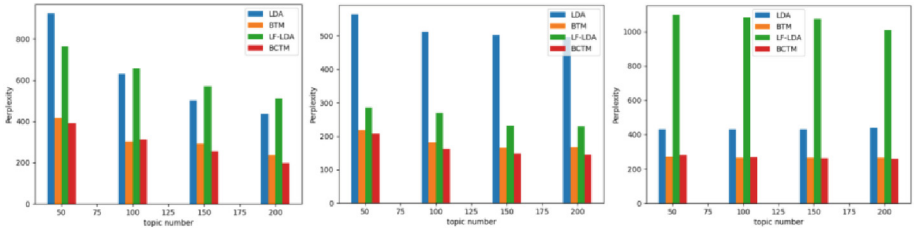

Fig. 2. Schematic diagram of confusion degree of WS, KOS and NIPS datasets

It can be concluded from the graph that due to the introduction of external information, the BCTM model not only expands the content of the text itself, but also improves the performance of confusion compared with the BTM model, and the introduction of word feature information is better for the LF-LDA model. If only in terms of the degree of confusion, the LDA model is the lowest, followed by LF-LDA, and then the best model of BTM, that is, the BCTM model proposed in this paper.

Next, we will evaluate the differences between the BCTM model and the other three models in terms of thematic consistency. Topic consistency calculates the semantic

consistency of words that belong to the same topic in the topic model. In the topic model, if the topic consistency is high, it shows that the words within the topic have similar semantics, and the classification effect of the topic model is good. If the consistency of the topic is low, it means that the internal expression of the topic is scattered, and the effect of topic classification is not good. This article uses Normalised Pointwise Mutual Information (NPMI) to measure the consistency of each topic in the topic model, and its formula is shown in formula 6.

$$\sum_{j=2}^{T} \sum_{i=1}^{j-1} log \frac{p(w_j, w_i)}{p(w_j)p(w_i)} / -logp(w_j, w_i) \tag{6}$$

Through this formula, the topic consistency score of topic k based on T high-frequency words in the topic can be calculated. P (w) is the probability of occurrence of word w, $p(w_i, w_j)$ is the probability of simultaneous occurrence of word w_i, w_j in a sliding window. In this experiment, the topic consistency score of each topic k is calculated by the participation of 10 high-frequency words of the topic.

In order to make the theme modeling effect more intuitive, the first ten words of the topic extracted from the BCTM model are mainly displayed. The model has a total of 50 topics, and the first five are shown here. The dataset modeled is KOS. The Table 4 shows.

The content of KOS dataset mainly comes from some topics and contents of blog, so its extracted topics are mostly related to people, events and so on. Take Topic1 as an example, the third word that appears frequently is Wayne, is the name of a character, while the Chinese meanings of the other words are “hand”, “drug management”, “management” and “debunk”. It can be seen from the subject words that the theme should be related to the illegal transactions carried out by a certain character. Topic2 is a theme-word distribution around the subject words such as “government”, “law enforcement” and “deadline”. Therefore, it can be inferred that the theme should be related to the implementation of laws and regulations of a certain country.

Table 4. TOP10 Vocabulary Presentation of KOS Dataset

Topic1	Topic2	Topic3	Topic4	Topic5
Millers	Grasp	Notably	Handful	Grasp
Hand	Enforcement	Kurds	Journal	Gains
Wayne	Browser	Spoke	Citation	Fundrasin
Drugs	Boulder	Posters	Address	Browser
Manage	Gains	Selection	Violent	Disgruntled
Debunking	Deadline	Implied	Enlisted	Manage
Fixed	Arrogant	Ball	Remained	Recount
Inform	Martin	Boost	Invited	Amendments
Offering	Latif	Judge	Succeed	Childhood
Page	Governmental	Powell	Gains	rallies

The calculation of topic consistency is inseparable from the topic-word matrix generated after the topic modeling, that is, the complete version of the previous Top- vocabulary. The word vector used to judge the contribution probability of topic words is obtained by pre-training the data set extracted from Wiki encyclopedia, and the size is about 5.48G. The specific experimental data are shown in Table 5.

Table 5. Comparison of Theme Consistency Among Models

Data set/$K = 50$	LDA	BTM	LF-LDA	BCTM
WS	−0.028	−0.144	−0.146	**0.023**
KOS	−0.083	−0.098	−0.078	**−0.053**
NIPS	−0.070	−0.087	**0.028**	−0.053
Data set/$K = 100$	LDA	BTM	LF-LDA	BCTM
WS	−0.036	−0.153	−0.152	**−0.001**
KOS	−0.089	−0.090	−0.084	**−0.053**
NIPS	−0.086	−0.002	**−0.010**	−0.046
Data set/$K = 150$	LDA	BTM	LF-LDA	BCTM
WS	−0.033	−0.149	−0.151	**−0.010**
KOS	−0.089	−0.088	−0.075	**−0.051**
NIPS	−0.085	−0.002	**−0.011**	−0.045
Data set/$K = 200$	LDA	BTM	LF-LFA	BCTM
WS	−0.138	−0.148	−0.114	**−0.004**
KOS	−0.089	−0.080	−0.076	**−0.046**
NIPS	−0.037	−0.006	**−0.002**	−0.035

The larger the value of the theme consistency is, the more relevant the subject words of the model are to the topic, and the most prominent experimental results in the table have been bolstered in boldface. As can be seen from Table 5, the BCTM model and the LF-LDA model are more effective in terms of thematic consistency. The topic consistency of the BCTM model is higher on WS datasets and KOS datasets with an average text vocabulary of 10 or 100. BCTM model not only uses word feature information to increase the relevance of subject words, but also introduces conceptual information to model the relationship between words, while BTM and LDA models either enrich text information, but do not enhance the correlation between words, and even introduce additional noise, so the topic consistency scores of these two models are relatively low. In some cases, the LDA model even performs better than the BTM model, which shows that the strong hypothesis of the BTM model does not necessarily play a positive role in the extraction of subject words.

On the whole, from the calculation of Top10 topic words to topic consistency, the comprehensive performance of BCTM model is better in short texts, and it is not weaker

than other models in medium-and long-term texts, which verifies the effectiveness of the introduction of external information on topic word extraction.

6 Conclusion

Topic model can effectively extract text features, and has been widely used and studied by industry and academia. With the rapid development of social networks, most of the texts in the network are gradually replaced by short texts, but due to the sparsity of the co-occurrence of short texts, the effect of traditional topic model modeling on short texts is not good. At present, most of the improvements to the short text topic model are carried out through document aggregation and the introduction of word vector features, but ignore high-quality prior knowledge such as conceptual information semantic network. This paper mainly studies the above problems, and completes the following work:

1) A short text topic model BCTM is constructed. BCTM model not only enriches the text content, but also introduces two kinds of external information, word feature information and concept information, into the topic model. At the document level, Bi-Concept sets are constructed for short text by introducing conceptual information, which expands the content of the text and improves the accuracy of modeling by the way of concept pair at the same time. At the modeling level, a special Dirichlet distribution is constructed by introducing word feature information as a priori knowledge, which makes the occurrence probability of similar words under the same topic more similar, thus improving the topic consistency.
2) In this paper, a method of transforming word vectors into binary tags is proposed, and the word feature information is introduced into the BCTM model. In this method, the prominent features of word vectors can be retained, while the weaker features can be discarded. At the same time, in order to apply the generated tags to the model, the topic-word feature tag correlation weight is calculated by introducing a new distribution, and then the influence on the Dirichlet prior parameters in the topic-word matrix is obtained. For each word under the theme, its unique Dirichlet prior parameters are calculated to achieve the purpose of improving topic consistency.
3) In this paper, a method of enriching short text information based on ConceptNet is proposed, and the conceptual information is introduced into the BCTM model. According to the algorithm proposed in this paper, the ConceptNet pair set is constructed according to the algorithm proposed in this paper, and the concept is introduced as external information to improve the accuracy of topic modeling. The BCTM model samples the set on the constructed concept. The follow-up is the parameter derivation of the above-mentioned method.

Although the research work of this paper has improved the quality of modeling in the topic model, the use of external information still needs to be explored. The concept nodes in ConceptNet usually have multiple parts of speech, but this paper takes the noun part of speech to construct Bi-Concept pairs. In the use of ConceptNet network, how to further refine the method of introducing conceptual information into topic modeling by means of part of speech tagging is the future research direction of this paper.

Acknowledgments. This work is supported by Key Research and Development Projects of Heilongjiang Province under grant number GA21C020, and Natural Science Foundation of Heilongjiang Province under grant number LH2021F015.

References

1. Ye, J., Zou, B., Hong, Y., Shen, L., Zhu, Q., Zhou, G.: Negation and speculation scope detection in Chinese. J. Comput. Res. Dev. **56**(7), 1506–1516 (2019). (in Chinese)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **34**(5), 993–1022 (2003)
3. Liu, Y., Wang, Z., Hou, Y., Yan, H.: A method of extracting malware features based on probabilistic topic model. J. Comput. Res. Dev. **56**(11), 2339–3234 (2019). (in Chinese)
4. Lee, Y.Y., Ke, H., Yen, T.Y., et al.: Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. J. Am. Soc. Inf. Sci. **71**(6), 657–670 (2020)
5. Limwattana, S., Prom-On, S.: Topic modeling enhancement using word embeddings. In: 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE) (2021)
6. Zhao, H., Du, L., Liu, G., et al.: Leveraging meta information in short text aggregation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
7. Yan, X., Guo, J., Lan, Y., et al.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456 (2013)
8. Wu, T., Qi, G., Wang, H., et al.: Cross-Lingual taxonomy alignment with bilingual biterm topic model. In: AAAI, pp. 287–293 (2016)
9. Zhu, Q., Feng, Z., Li, X.: GraphBTM: graph enhanced autoencoded variational inference for biterm topic model. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4663–4672 (2018)
10. Li, X., Zhang, A., Li, C., et al.: Relational biterm topic model: Short-text topic modeling using word embeddings. Comput. J. **62**(3), 359–372 (2019)
11. Huang, J., Peng, M., Li, P., et al.: Improving biterm topic model with word embeddings. World Wide Web **23**(6), 3099–3124 (2020)
12. Nguyen, D.Q., Billingsley, R., Du, L., et al.: Improving topic models with latent feature word representations. Trans. Assoc. Comput. Linguist. **3**, 299–313 (2015)
13. Li, C., Wang, H., Zhang, Z., et al.: Topic modeling for short texts with auxiliary word embeddings. In: International ACM SIGIR Conference, pp. 165–174. ACM (2016)
14. Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G.: Incorporating word embeddings into topic modeling of short text. Knowl. Inf. Syst. **61**(2), 1123–1145 (2018). https://doi.org/10.1007/s10115-018-1314-7
15. Yi, F., Jiang, B., Wu, J.: Topic modeling for short texts via word embedding and document correlation. IEEE Access **PP**(99), 1 (2020)
16. Archambeau, C., Lakshminarayanan, B., Bouchard, G.: Latent IBP compound Dirichlet allocation. IEEE Trans. Pattern Anal. Mach. Intell. **37**(2), 321–333 (2014)
17. Wu, X., Li, C., Zhu, Y., et al.: Short text topic modeling with topic distribution quantization and negative sampling decoder. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
18. Wallach, H.M., Minmo, D., Mccallum, A.: Rethinking LDA: why priors matter. Adv. Neural. Inf. Process. Syst. **23**, 1973–1981 (2009)
19. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: The 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–533 (2014)