Enhancing Short Text Topic Modeling with FastText Embeddings

Fan Zhang[1]
College of Computer Science,
Wuhan Donghu University
Wuhan, China
e-mail: whzhangfan@126.com
Fan Zhang and Wang Gao contribute equally to the article

Wang Gao[2, *]
School of Mathematics and Computer Science,
Jianghan University
Wuhan, China
e-mail: gaow@jhun.edu.cn
Fan Zhang and Wang Gao contribute equally to the article

Yuan Fang[3]
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: fangyuan2000@foxmail.com

Bo Zhang[4]
School of Computer Science
Wuhan Donghu University
Wuhan, China
e-mail: bob.cheung@ovspark.com

***Abstract*—Over the past few years, we have experienced the rapid development of online social media, which produced a variety of short texts. It is important to understand the topic patterns of these short texts. Because of data sparsity, traditional topic models are not suitable for short text topic analysis. In this paper, we proposed a novel topic model, referred as FastText-based Sentence-LDA (FSL) model, which extends the Sentence-LDA topic model for short texts. We first utilize the FastText model to train a word embedding replacement model, which can alleviate the problem of lacking word co-occurrence information over short texts. Secondly, we propose a new latent feature topic model which integrates latent feature word embeddings into Sentence-LDA. Experimental results demonstrate that our new model has produced significant improvements in topic coherence by using information from external corpora.**

Keywords- Enhancing short text topic modeling, Sentence-LDA Topic Model, Word Embedding Replacement Model

I. INTRODUCTION

With the advent of a large number of text corpora from various sources on the Internet, such as blogs, social networks and news media, it is becoming more and more important to explore a high-level understanding of these texts. Conventional topic models assume that documents are generated from a mixture of shared topics at the corpus level, such as Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA). However, for scarcity statistics in short texts, the assumption of a topic mixture for each document is considered too rich.

Many heuristic strategies have been adopted to alleviate the sparsity problem for topic modeling over short texts. Li et al. proposed a topic model that promotes the semantically related words to share the same topic label during the sampling process using a Generalized Pólya Urn (GPU) model [1]. Gao et al. proposed a generalized method for short text topic modeling by using the Embedding-based Minimum Average Distance (EMAD) [3]. Bunk proposed a topic model that combines word embeddings with LDA to enhance topic quality [3]. Based on the current research, the existing short text topic modeling methods have the following two shortcomings: (1) Traditional methods improve the semantic information of a corpus by using an external corpus to expand the meaning of words or merge short texts. However, the topic model does not adequately extract the meaning information of the training corpus. (2) The efficiency of the word embedding based topic model is very slow when it runs in the inner layer of Gibbs sampling. In this paper, we propose a novel topic model of short texts, referred as FastText-based Sentence-LDA (FSL) model. The contributions of this paper are summarized as follows: (1) FSL employs FastText to train word w' that replaces word w in the Sentence-LDA model, which can effectively alleviate the sparsity problem. (2) FSL integrates word vector representations into Sentence-LDA using a Word Embedding Replacement Model (WERM). To the best of our knowledge, this is the first work of Sentence-LDA, combining word embeddings with WERM model. (3) Experimental results demonstrate FSL outperforms baseline models on topic coherence.

II. RELATED WORK

A. Sentence-LDA Topic Model

The Sentence-LDA model proposed by Jo et al. is an extension of the LDA model [4]. The main idea of Sentence-LDA is that all words in a short text are generated from one topic. This simplified method is not suitable for lengthy documents, but it can alleviate the sparsity problem for short texts. Therefore, fine-grained semantic information extraction can be achieved by analyzing the structure of sentences or phrases.

978-1-7281-6499-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICBAIE49996.2020.00060

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on June 09,2021 at 13:19:06 UTC from IEEE Xplore. Restrictions apply.

Based on LDA's three-tier structure of "text-topic-words", Sentence-LDA adds a sentence layer between text and topic layer, which becomes a four-tier structure of "text-sentence topic-words". Figure 1 shows the graphical representation of Sentence-LDA.

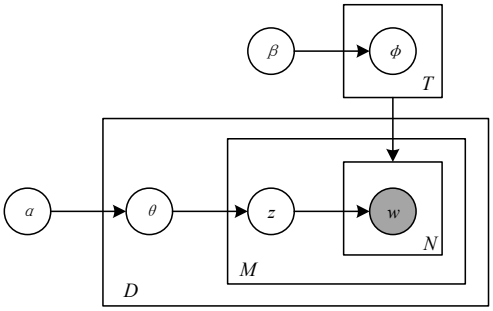

Figure 1. The graphical representation of Sentence-LDA.

Sentence-LDA is a generative model, and its generation process is as follows:

- For each topic label z, draw a topic-word distribution $\phi_z \sim \text{Dirichlet}(\beta)$.
- For each short text d,
  - draw a document-topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$,
  - draw a topic $z \sim \text{Multinomial}(\theta_d)$.
  - For each word w, draw a word $w \sim \text{Multinomial}(\phi_z)$.

B. FastText Word embedding

Word embeddings proposed by Mikolov et al. [5] have been successfully applied to several text mining tasks such as emerging topic tracking and text compression [6,7]. In this paper, we exploit FastText that is a lightweight, free, open-source library, to learn word embeddings. FastText can efficiently train on large-scale datasets, and the word embeddings learned by this tool are able to capture the similarity between words.

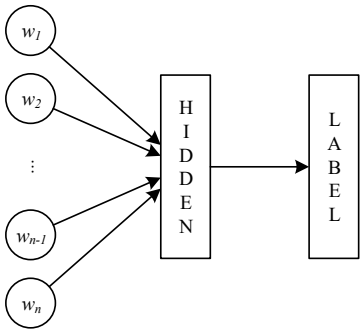

Figure 2. The model architecture of FastText.

A simple FastText model structure with a hidden layer is shown in Figure 2. The model uses a string of words as input and generates a probability distribution on predefined classes. The FastText word embeddings are averaged as a text representation, which is fed to a linear classifier.

III. METHODOLOGY

A. Word Embedding Replacement Model

The semantic relations of words have similarities and correlations. For example, words "person" and "individual" have a semantic similarity relation, while words "teacher" and "student" have a semantic correlation relation. Models based on word embedding focus on semantic similarity, while document-based topic models are good at capturing semantic relevance. Considering that the amount of datasets in the experiment is not very large, the FastText model is used to construct the word embedding model.

The corpus is trained by the FastText model, which can well represent similar words. The cosine distance is used to calculate the similarity between words. We define the cosine distance between w_i and w_j as

$$d(w_i, w_j) = 1 - \frac{v_i \bullet v_j}{\|v_i\| \bullet \|v_j\|}, \tag{1}$$

where v_i and v_j are word vectors corresponding to word w_i and w_j respectively.

TABLE I. EXAMPLES OF THE MOST SIMILAR WORD EMBEDDING COSINE DISTANCE OF "TEACH"

Word	*Cosine Distance*
learn	0.7246
taught	0.7122
instruct	0.7081
educate	0.6871
teaches	0.6736

Table 1 shows the words with similar meanings of "teach" that were captured after entering the corpus under the FastText model in the experiment.

In the word embedding replacement model, the specific method of replacing word w is to extract a word w' from the FastText model. w' is a word with the cosine distance closest to w in the word embedding space. For instance, as shown in Table 1, the FSL model employs word "learn" to replace "teach".

Follow the LF-LDA model [8], we introduce a Bernoulli parameter $\eta \sim \text{Bernoulli}(\varepsilon)$ to construct the word embedding replacement model. The sampling of words can be generated from the word embedding space or from the word distribution ϕ_z with a certain probability.

When the parameter ε is set to 0, the model degenerates into a traditional Sentence-LDA, and words are generated directly from the topic distribution. When the parameter ε is not equal to 0, the model extracts new words from the word embedding space with a certain probability. The parameter ε

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on June 09,2021 at 13:19:06 UTC from IEEE Xplore. Restrictions apply.

determines the extent to which additional information of words leaks into the process of topic inference.

B. FSL Model

In the FSL model, the preprocessed text is first input into the FastText model to obtain a trained word embedding space. Secondly, a word embedding replacement layer is added to the FSL model. Finally, word w is input into the word embedding replacement layer. The structure of the model is shown in Figure 3.

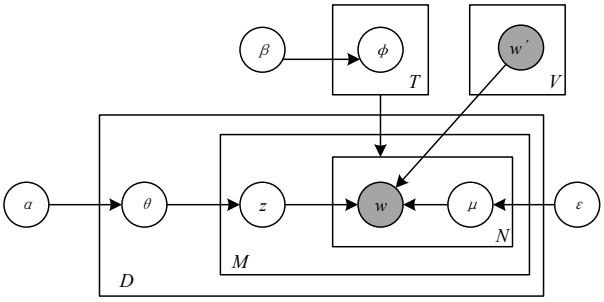

Figure 3. The graphical representation of FSL

The generation process of FSL is as follows:

- For each topic label z, draw a topic-word distribution $\phi_z \sim \text{Dirichlet}(\beta)$.
- For each short text d,
  - draw a document-topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$,
  - draw a topic $z \sim \text{Multinomial}(\theta_d)$,
  - draw a Bernoulli parameter $\eta \sim \text{Bernoulli}(\varepsilon)$,
  - for each word w, draw a word $w \sim \text{Multinomial}(\phi_z)$,
  - if $\eta = 1$, replace w with w'.

Word w' is the most similar word with w, which is captured by the word embedding replacement model. Therefore, the topic label of the i_{th} word is drawn from the following conditional probability:

$$P(z_i = k \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{C_{dk}^{DT} + \alpha_k}{\sum_{k'}^{T} C_{dk'}^{DT} + \alpha_{k'}} \frac{\Gamma(\sum_{w=1}^{M} C_{kw}^{TM} + \beta_w)}{\Gamma(\sum_{w=1}^{M} (C_{kw}^{TM} + \beta_w) + m_i)} \prod_{w=1}^{M} \frac{\Gamma(C_{kw}^{TM} + \beta_w + m_{iw})}{\Gamma(C_{kw}^{TM} + \beta_w)}, \quad (2)$$

where D, T and M are the number of short texts, topics and words, respectively. C_{dk}^{DT} denotes the number of short texts that are assigned to topic k, and C_{kw}^{TM} denotes the number of words that are assigned to topic k. $m_{i(w)}$ is the number of total words (or word w) in short text i. Based on the Bernoulli distribution, word w' is sampled from the word embedding replacement model layer, and the distribution of the topic of the current word w is exchanged. Since the training of word embedding does not run in the layer of Gibbs sampling, but after the FastText model is trained, FSL replaces words with similar meanings extracted from the embedding space with a certain probability in the word sampling stage.

The generation probability of topic k in short text d is

$$\theta_{dk} = \frac{C_{dk}^{DT} + \alpha_k}{\sum_{k'=1}^{T} C_{dk'}^{DT} + \alpha_{k'}}. \quad (3)$$

The generation probability of word w in topic k is

$$\phi_{kw} = \frac{C_{kw}^{TM} + \beta_k}{\sum_{w'=1}^{V} C_{kw}^{TM} + \beta_k}. \quad (4)$$

Word substitution can alleviate the sparseness problem of short text topic modeling, and training word embedding space externally can improve the efficiency of the proposed model.

IV. Experiment

In this section, we report the experimental results to demonstrate the effectiveness of our new topic model by comparing it with state-of-the-art baselines on two real-world corpora of short texts.

A. Dataset and Setting

We use two datasets in the experiments: Questions: the corpus contains 10512 questions from Quora[1], a well-known online Q&A community. The average length of a question is 11.53 words, which is typical short text. News is a dataset of 7,290 English news titles crawled from reuters.com. The average length of each title is 15.42 words. We perform the following preprocessing step on both datasets: (1) delete stop words and non-alphabetic characters; (2) delete words with document frequency less than 3.

We compare our model with two baseline models: LF-LDA replaces LDA's topic-to-word Dirichlet multinomial distribution with a two-component mixture of a topic-to-word Dirichlet multinomial distribution and a potential feature distribution [8]. Gaussian LDA replaces LDA's "topic" parameterization with multivariate Gaussian distribution in the embedding space [9].

For both baselines, we implement two models using the tools released by the authors, and choose the parameters according to their original papers. For all the methods in comparison, we empirically set $\alpha = 50/N$, $\beta = 0.01$, $\varepsilon = 0.5$, where N is the number of news titles. In the experiment, we use pre-trained FastText word vectors[2] learned on UMBC webbase corpus, statmt.org news and Wikipedia 2017 dataset. If a word does not have a word embedding, it is assumed that the word will not be exchanged in the process of topic inference.

[1] http://www.quora.com/

[2] http://fasttext.cc/docs/en/english-vectors.html

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on June 09,2021 at 13:19:06 UTC from IEEE Xplore. Restrictions apply.

B. Evaluation Measures

In this study, we use UCI topic coherence to evaluate the effectiveness of topic models. UCI introduced in [10] exploits Point-based Mutual Information (PMI) to measure topic coherence. A higher UCI value denotes that the topics generated by the model are more semantically coherent. Given a topic k represented by its top- N words w_1, w_2, ..., w_N, the UCI value for k is:

$$UCI(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_a, w_b), \tag{5}$$

$$PMI(w_a, w_b) = \log \frac{p(w_a, w_b)}{p(w_a)p(w_b)}, \tag{6}$$

where $p(w_a)$ and $p(w_b)$ denote the probabilities of words w_a and w_b appearing in a sliding window, respectively. $p(w_a, w_b)$ denotes the probabilities of words w_a and w_b co-occurring in a sliding window. The UCI score of a topic model refers to the average score of all topics generated by the model.

C. Experimental Results

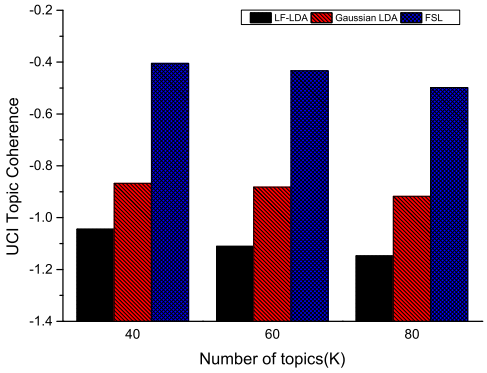

Figure 4. Topic coherence on questions dataset.

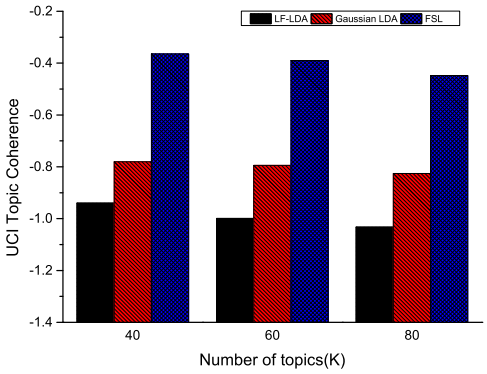

Figure 5. Topic coherence on news dataset.

Figure 4 and 5 illustrate the UCI topic coherence on the two datasets with number of top words per topic T = 10 and number of topics ranging from 40 to 80, respectively. The experimental results prove that the proposed model significantly performs better than LF-LDA and Gaussian LDA (P-value < 0.01 by t-test). It is reasonable that FSL is superior to LF-LDA on such two datasets because LDA lack sufficient word information in the case of short text and cannot reveal the semantic relationship at the word-level. Modeling the corpus based on word-level relationships can enhance the quality of topics, and thus the performance of Gaussian LDA is better than LF-LDA. In addition to explicitly modeling the relationship between words, FSL also finds semantic relationships between pairs of non-co-occurring words at the corpus level.

V. Conclusion

In this paper, we propose a new topic model for short texts, namely FastText-based Sentence-LDA (FSL) model. FSL first exploits the FastText model to train word embedding replacement model, which can effectively alleviate the sparsity. Furthermore, our model integrates latent feature word embeddings into Sentence-LDA to improve topic modeling over short texts. We conducted experiments on two real-world short text datasets, i.e. question and news title corpora. The experiment results demonstrate the effectiveness of our model compared with two state-of-the-art baselines. In the future, we intend to adjust the Bernoulli parameter ε based on the related word pairs, which is a fixed value in this paper. In addition, we would like to reveal whether FLS can also work well on lengthy document datasets.

Acknowledgements

This paper was cosupported by the Foundation for Young Scholars in Wuhan Donghu University under grant 2018dhzk008. Thanks for the support of Wuhan Donghu University. Thank you for providing the experimental platform in Wuhan Donghu University. We have completed the implementation of the algorithm on this platform and the experimental data is presented.

References

[1] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun and Zongyang Ma. Topic Modeling for Short Texts with Auxiliary Word Embeddings. Proceeding of the International Conference on Research and Development in Information Retrieval (SIGIR), 2016, pp:165-174.

[2] Wang Gao, Min Peng, Hua Wang, Yanchun Zhang, Qianqian Xie and Gang Tian. Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems, 2018, (12):1-23.

[3] Stefan Bunk and Ralf Krestel. WELDA: Enhancing Topic Models by Incorporating Local Word Context. Proceedings of The 18th ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2018, pp:78-86.

[4] Yohan Jo and Oh Alice. Aspect and sentiment unification model for online review analysis. Proceedings of the 4th International Conference on Web Search and Web Data Mining (WSDM), 2011, pp:815-824.

[5] Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies (HIT-NAACL), pp:746-751.

[6] Jiajia Huang, Min Peng, Hua Wang, Jinli Cao, Wang Gao, Xiuzhen Zhang. A probabilistic method for emerging topic tracking in microblog stream. World Wide Web Journal, 20(2):325-350.

[7] Min Peng, Wang Gao, Hua Wang, Yanchun Zhang, Jiajia Hunag Qianqian Xie, Gang Hu and Gang Tian. Parallelization of massive

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on June 09,2021 at 13:19:06 UTC from IEEE Xplore. Restrictions apply.

textstream compression based on compressed sensing. ACM Transactions on Information Systems (TOIS), 2017, 36(2):1-18.

[8] Dat Quoc Nguyen, Richard Billingsley, Lan Du and Mark Johnson. Improving topic models with latent feature word representation. Transactions of the Association for Computational Linguistics (TACL), 2015, 3:299-313.

[9] Rajarshi Das, Manzil Zaheer and Chris Dyer. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), 2015, pp:795-804.

[10] David Newman, Jey Han Lau, Karl Grieser and Timothy Baldwin. Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), 2010, pp: 100-108.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on June 09,2021 at 13:19:06 UTC from IEEE Xplore. Restrictions apply.