ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 392-397
doi: 10.18178/wcse.2017.06.067

Enhancing Topic Models for Short Texts using Deep Autoencoder

Luepol Pipanmaekaporn[1] and Suwatchai Kamolsantiroj[2]

[1,2] Department of Computer and Information Science,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand 10800

Abstract. With the prevalence of micro-blogging platforms such as Twitter, short texts are becoming a large portion of online text data. Despite this, inferring meaningful topics from short texts is still a challenging task due to very limited word co-occurrence information available in documents. In this paper, we propose a novel way for topic learning and inference in short text data by exploring semantic relations between words over short texts and then building word relation matrix to avoid the data sparsity issue. Deep Autoencoder is used to learn meaningful relations between words and topics by using the auxiliary information of word relations. After acquiring the topics, we formulated the problem of topic inference as Non-Negative Matrix Factorization (NNMF) with word-topic relations. Experiments on two real-world short text datasets show that our method outperforms state-of-the-art baselines for topic modeling.

Keywords: Short Text, Topic Modeling, Short Text, Representation Learning, Deep Autoencoder

1. Introduction

Topic modelling has been proven to be useful for automatic topic discovery from a huge volume of texts. Topic model views texts as a mixture of latent topics, where a topic can be represented by either non-probabilistic or probabilistic distributions over words. For example, one of the most typical probabilistic topic model, namely Latent Dirichlet Allocation (LDA) [1], has achieved great success in modelling text collections. Alternatively, Non-Negative Matrix Factorization (NNMF) is very popular for a variety of text processing tasks based on matrix factorization [2]. Although several experimental results have shown that topic models are effective for dealing with the data sparsity of texts, the direct use of conventional topic models can be affected by very limited word co-occurrence information in short text. This is because the difficulty in acquiring senses of words. To alleviate the sparsity of short documents, a great amount of efforts have been made by enriching short text information using auxiliary information of external knowledge bases such as Wikipedia [3][5]. The major disadvantage of these works is that the auxiliary information is not always available and does not take features of short texts into account. Some of the efforts have been attempted intensifying word co-occurrence information in short text to improve conventional topic models. For example, biterm topic model (BTM) [4] that generates word co-occurrence patterns instead of single words. Recently, a strategy of aggregating short texts into regular-sized pseudo documents before performing a standard topic model was widely applied [6].

Motivated by these works, we focus on exploring semantic relations between words as auxiliary information for topic learning in short texts without the consultant of external information. We first build word relation matrix based on the semantic similarity of words. To do this, we apply skip-gram model of word2vec [7], to acquire distributed vector representation of words which captures the semantic and syntactic relationships. After obtaining the word relations, Deep Autoencoder [8] is used to learn meaningful relations between words and (latent) topics. We finally formulate the problem of topic inference as Non-

[+] Corresponding author.
E-mail: luepol.p@sci.kmutnb.ac.th[1] and suwatchai.k@sci.kmutnb.ac.th

Negative Matrix Factorization (NNMF) with auxiliary information of the word-topic relations learned by neural network. Figure 1 illustrates the entire process of the proposed algorithm.

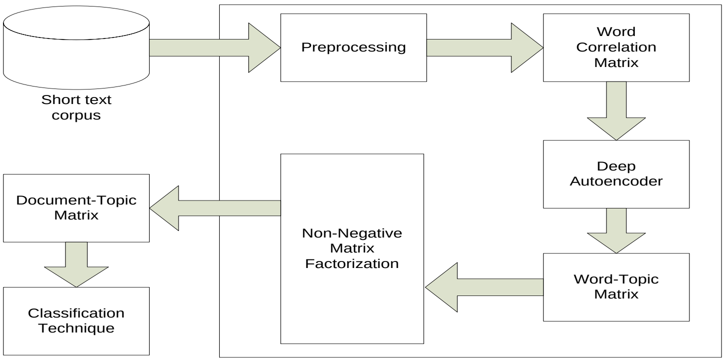


Fig. 1: The proposed framework.

2. Our Approach

2.1. Word relation matrix

Conventional topic models, such as LDA and NNMF, typically learn topics over corpus from word-document matrix. However, the word-document matrix that represents information about word occurrences in documents is extremely sparse in short texts. As a result, these existing methods often suffer from the data sparsity. Instead of word-document matrix, we explore semantic correlations between words as auxiliary information for topic discovery in short text corpus. The main advantage of using this information is that the word relation data is dense and captures correlations between words that are important for topic learning. Furthermore, when collection size of documents increases, the number of words tends to be a stable value. Consequently, it is suitable for large-scale document collections.

To build the word correlation matrix, we apply word2vec, an efficient neural probabilistic language model from Google [7], that learns distributed vector representation of words from plain texts. The feature vectors capture syntactic and semantic regularities of words within a window size. In this work, we choose continuous skip-gram model, one of the word2vec models, to learn the feature vectors for words over corpus. The skip-gram model learns the probability of context words given a word $\mathrm{w_t}$ by minimizing the loss function: $\mathrm{E} = -\log \mathrm{p}(\mathrm{w_{t-j}}, \dots, \mathrm{w_{t+j}} | \mathrm{w_t})$ where j indicates a window size. We then segment the pre-processed short texts into a collection of sentences as input for the skip-gram model. Table 1 demonstrates parameter settings used for Word2vec. After learning the feature word vectors, cosine coefficient similarity measure is applied to compute the correlation between any two words, resulting in the final word correlation matrix.

2.2. Topic learning and inference

In this section, we will explain the details of our algorithm for topic discovery from the word correlation information. After we build word correlation matrix X, we apply autoencoder (AE) [9], a neural network that learns compact and high-level feature representation of input data which is originally of high-dimension. AE consists of two main parts, namely encoder and decoder parts. The encoder part will represent the high-dimensional data into low-dimension while the decoder will convert the low-dimensional representation into high-dimensional output. Mathematically, AE learns the output $\hat{X} = \{\hat{x}_1, \hat{x}_2, .., \hat{x}_n\}$ to reconstruct the input $X = \{x_1, x_2, .., x_n\}$ with a low-dimensional hidden layer $Z = \{z_1, z_2, .., z_k\}$ where $0 < k < n$. The value of hidden layer node z_i is defined as

$$z_i = f(b_i^{(1)} + \sum_{j=1}^{n} w_{ij}^{(1)} x_j) \tag{1}$$

where z_i is the ith hidden layer unit. $\mathrm{b_i^{(1)}}$ and $\mathrm{w_{ij}^{(1)}}$ indicate the bias of hidden node i and the weight between input node j and hidden node i. f is a nonlinear activation function. At the decoder stage, the output layer $\hat{\mathrm{X}}$ is built with the activations of the hidden layer as input, bias $\mathrm{b}^{(2)}$ and eights $\mathrm{W}^{(2)} = \left\{\mathrm{w}_{11}^{(2)}, \mathrm{w}_{12}^{(2)}, \dots, \mathrm{w}_{\mathrm{kn}}^{(2)}\right\}$:

$$\hat{x}_i = f\left(b_i^{(2)} + \sum_{j=1}^{k} w_{ij}^{(2)}\, a_j\right) \qquad (2)$$

where a_j is the output activation of j^th hidden node. The network is trained to minimize the errors between input vector X and output vector $\hat{X}$. AE can be recently made deep by adding more encoding and decoding layers, known as deep autoencoder (DAE) [8]. In this work, we use DAE to learn meaningful topic representation of words from word-word matrix. Figure 2 shows our DAE architecture used for topic learning.

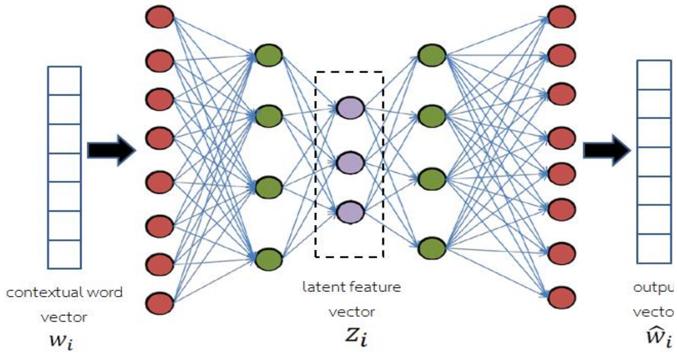


Fig. 2: Deep Autoencoder architecture for topic extraction

According to Figure 2, the architecture of the deep autoencoder consists of five layers: input layer, output layer, two hidden layers for encoding and a hidden layer for decoding. All units have sigmoid as nonlinearity in all the layers. The network is trained by means of unsupervised pre-training of the deep structure [13] where each subsequent layer was trained by using data which were derived from the previous trained layer. Fine-tuning of the whole deep structure was performed by using mini-batch gradient descent with mean square error (MSE) cost function. The weight decay parameter of λ is used to control the relative importance of the regularization calculated as $||W||^2$. More details regarding training of a deep autoencoder can be referred to the work [10]. After training the network, we extract the hidden units in latent feature layer for word-topic matrix Z. After acquiring the distribution of words under the topics, we formulate this topic learning problem as Non-Negative Matrix Factorization (NNMF), which estimates topic-document matrix U by minimizing the following objective function:

$$D(U) = ||X - ZU||_F^2 \qquad (3)$$

where X, Z denote the word-document matrix and the word-topic matrix respectively. As seen in Eq. (3), the topic-document matrix U can be estimated by minimizing the Euclidean distance between the term-document matrix X and the product ZU, given the topic-document matrix U. To efficiently solve the problem, we use non-negative least square [2] to yield an approximate solution by enforcing all the negative elements to zero:

$$\hat{U} \leftarrow max\ ((Z^T Z)^{-1} Z^T X, 0) \qquad (4)$$

3. Experiments

In this section, we report empirical experiments to evaluate our proposed method, named DAE, on two common tasks of short text classification: 1) Tweet sentiment analysis and (2) News title classification.

3.1. Datasets and Experimental Settings

We also demonstrate the effectiveness of DAE on two real-world datasets: Sentiment140 [2] and 20newsgroup[3] and compare it with the conventional topic models: LDA and NNMF. We would like also to compare DAE with a standard autoencoder with a single hidden layer, named AE. The twitter dataset contains 1,600,000 English tweets collected from Twitter during April, 6, 2009 to June 25, 2009. Each tweet

[2] http://www.sentiment140.com

[3] http://www.ai.mit.edu.edu/~jrennie/20Newsgroups

was hand-classified with either positive or negative sentiment. All special characters and emoticons are removed. We select only 4,000 tweets in each sentiment category for the evaluation of effectiveness. For news title classification, we choose the well-known 20 newsgroups dataset, which contains over 20,000 English posts from 20 different newsgroups. We only experiment with six of these groups, including comp.graphics, rec.autos, rec.sport.baseball, sci.electronics, soc.religion.christian and talk.politics.guns. For the short text scenario, we only consider the subject filed of each document as training and test data and discard other information such as main body. Table 2 shows statistics of the two datasets. We evaluate all the methods using a linear support vector machine classifier LIBLINEAR [11], a software library for large-scale linear classification. In each dataset, documents are randomly split into training, validation and testing with 70%, 10% and 20% ratio respectively. For multi-class classification, we use the one-vs-the rest approach to train the SVM classifier.

Table 1: Statistical Information of the two datasets

Dataset	Twitter	news title
#documents	80,000	5,845
#words	2,514	1,860
average words per document	12.59	9.26
#classes	2	6

Our proposed method consists of two major steps, including building word correlation matrix using skip-gram model of word2vec and topic learning using deep autoencoder (DAE). To train the skip-gram model, we used a minimum count of four word occurrences and used window size 5. The dimensionality of the word vectors is equal to 200 because this value performs the best performance for the datasets. The deep architecture of neural network consists of five layers: input – encode – topic – decode – output layers. 500 units are set in the layers of encoding and decoding. We also use the validation set to find the best parameters by training the SVM classifier on the validation sets of each dataset and varying the number of k units in topic layer from 20 to 80 and the value λ from 0.001 to 0.09. The number of iterations in NNMF is set 1,000.

3.2. Experimental Results

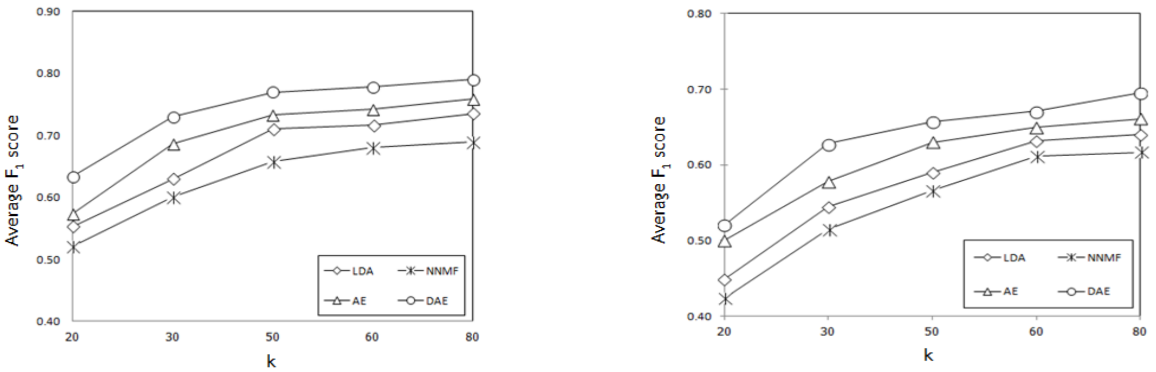
To evaluate the performance, the common measure F_1 score was used. Figure 3 includes the average F_1 score results for each of the methods performed on the two datasets. As can be seen from Figure 3, SVM generated the best F_1 scores with our method (DAE) at all the k values where it reached the best performance on 20newsgroups dataset at $\mathrm{k} = 80, \lambda = 0.07$ and also on twitter dataset at $\mathrm{k} = 80,\ \lambda = 0.04$. SVMs generated by LDA and NNMF were also achieved the best performance at between $\mathrm{k} = 40$ and $\mathrm{k} = 80$ for both the datasets. The most interesting findings revealed in this table is that both DAE and AE that learn topics with auxiliary information of semantic relations of words outperform LDA and NNMF models that use word-document relations, which is sensitive to over-fitting. The result supports the superiority of using information of word relations to alleviate the extreme sparsity of short documents. Furthermore, it highlights the use of autoencoder for learning meaningful relations between words and topics. As seen in Figure 3, the encouraging improvement of DAE that learns topics with the deep architecture of autoencoder is effective and consistent compared to AE that uses a shallow structure of autoencoder. This result confirms that the deep architecture is effective for topic discovery from words.

3.3. Experimental Results

To evaluate the performance, the common measure F_1 score was used. The F_1 measure is a combination of both precision and recall performance metrics that indicate the extent to which a group of classified documents by system belonging to a particular class. Figure 3 includes the average F_1 score results for each of the methods performed on the two datasets. As can be seen from Figure 3, SVM generated the best F_1 scores with our method (DAE) at all the k values where it reached the best performance on 20newsgroups dataset at $\mathrm{k} = 80, \lambda = 0.07$ and also on twitter dataset at $\mathrm{k} = 80,\ \lambda = 0.04$. SVMs generated by LDA and NNMF were also achieved the best performance at between $\mathrm{k} = 40$ and $\mathrm{k} = 80$ for both the datasets. The most interesting findings revealed in this table is that both DAE and AE that learn topics with auxiliary information of semantic relations of words outperform LDA and NNMF models that use word-document relations, which is sensitive to over-fitting. The result supports the superiority of using information of word relations to alleviate the extreme sparsity of short documents. Furthermore, it highlights the use of autoencoder for learning meaningful relations between words and topics. As seen in Figure 3, the encouraging improvemen t of DAE that learns topics with the deep architecture of autoencoder is effective and consistent compared to AE that uses a shallow structure of autoencoder. This result confirms that the deep architecture is effective for topic discovery from words.

4. Conclusion

Conventional topic models have very limited use in short texts due to the extreme sparsity of word co-occurrence patterns in documents. In this paper, we have proposed a novel method that combines word relation information and deep autoencoder to tackle this problem. Specifically, our method learns meaningful word-topic relation from word relation data using deep structure of autoencoder. After that, Non-Negative Matrix Factorization (NNMF) is applied to infer topic-document relation. The experimental results conducted on two benchmark datasets demonstrate that the proposed algorithm achieves performance improvements compared to state-of-the-art baselines, including LDA and NNMF. Future work is mainly focused on other applications such as clustering microblogs and hot topic detection in Twitter. We believe that the proposed algorithm is promising for dealing with the severe sparsity of short text data.



(a) Twitter Sentiment dataset

(b) News title dataset

Fig. 3: The average $\mathbf{F_1}$ scores w.r.t. the number of $\boldsymbol{k}$ topics on both the datasets.

5. Acknowledgements

This research was funded by King Mongkut's University of Technology North Bangkok. Contact no. KMUTNB-GEN-59-42.

6. References

[1] Blei, D.M., Ng, A.Y., Jordan, M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2003, 3(1): 993-1022.

[2] Lee, D. and Seung H. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* 2001. pp. 556-562.

[3] Phan, X.H., Nguyen L., Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proc. of 17th ACM international conference on World Wide Web*. 2008, pp.91-100.

[4] Yan, X., Guo, J., Lan, Y. and Cheng, X.: A biterm topic model for short texts. In *Proc. of 22nd international conference on World Wide Web*, 2013. pp.1445-1456.

[5] Li, C., Wang, H., Zhang, Z., Sun, A. and Ma, Z. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *Proc. of 39th International ACM SIGIR conference on Research and Development in Information Retrieval* 2016. pp. 165-174.

[6] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., Xiong, H. Topic Modeling of Short Texts: A Pseudo-Document View. In *Proc. of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 2105-2114.

[7] Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space. In *Proc. of 27th Annual Conference on Neural Information Processing Systems*. 2013, pp.3111-3119.

[8] Gehring, J., Miao, Y., Metze, F., Waibel, A. Extracting deep bottleneck features using stacked auto-encoders. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 3377-3381.

[9] Hinton, G. E. and Ruslan, R. S. Reducing the dimensionality of data with neural networks. *Science* 2006, **313**(5786): 504-507.

[10] Yu, D. and Seltzer, M.L. Improved Bottleneck Features Using Pretrained Deep Neural Networks. *Interspeech* 2011, **237**(1): 237-240.

[11] Joachims, T. Making large scale SVM learning practical. *Universität Dortmund* 1999.