A Method based on Sentence Embeddings for the Sub-Topics Detection

To cite this article: Yu Xie et al 2019 J. Phys.: Conf. Ser. 1168 052004

View the article online for updates and enhancements.
A Method based on Sentence Embeddings for the Sub-Topics Detection

Yu Xie\textsuperscript{1,2}, Bin Zhou\textsuperscript{1,2,a} and Yang Ou\textsuperscript{3}

\textsuperscript{1}National University of Defense Technology, Department of Computer, Hunan 410005, China
\textsuperscript{2}Laboratory of Software Engineering for Complex Systems, Hunan 410005, China
\textsuperscript{3}China Satellite Launch and Tracking Control General, Beijing 100720, China
\textsuperscript{a} Corresponding author. E-mail address: binzhou@nudt.edu.cn

Abstract. Weibo topics are characterized by diversity and complexity, and often contain different sub-topics even for the same topic. Therefore, how to classify sub-topics effectively and accurately is of great significance. Due to the strong similarity between sub-topics belonging to the same topic, most existing methods cannot be directly applied to the task of sub-topic discovery. In this paper, a new technique based on sentence embeddings for detecting sub-topics of weibo is proposed. The information of blogs is represented by sentence embeddings containing semantic information and the results are verified by clustering.

1. Introduction
Weibo is a popular social media. Usually, some hot events cause countless discussions on weibo. These events have the characteristics of long duration and wide coverage, and are accompanied by different sub-topics to describe different aspects of the event. For example, the events related to vaccine failure include: event description, reseeded vaccine scheme, social response and accountability, etc. These sub-topics not only describe the different perspectives of the event, but also reflect the possible focus of public opinion [1], which are of great help to public opinion analysis.

The traditional method of topic detection can be summarized as clustering of data sets, and each cluster can describe one topic. The main problem of traditional methods in finding sub-topics is that blogs describing the same event tend to be very similar, and the distinction of sub-topics is not very large, so the effect of clustering or classification is greatly affected.

This problem can be partly solved by the sentence vector. The sentence vector can be used as a dense, low-dimensional real value vector of the short text, which can capture the syntax and semantics of the short sentence. Consider that in the sub-topics detection, the same word may have different meanings under different sub-topics. To enhance the differentiation, our basic idea is allow each word have different vector representations under different topics.

Considering the above issues, this paper proposes a sub-topics detection method based on TWE model [6]. Our method uses the TWE model to train the weibo data set under a topic. The word embeddings and topic embeddings obtained are connected through the p-means method [9], and the sentence embeddings that retain semantic features are obtained. Then, sentence embeddings are used as text features for clustering to get sub-topics.

The main contribution of this paper is: (1) combine topic embeddings to generate sentence vectors, and enhance the semantic features of sentence embeddings. (2) proposed a sub-topics detection
method based on sentence embeddings to detect potential sub-topics under the same topic. (3) the model is applied to the weibo topic set about public events, which proves that its effect is better than the existing topic model.

2. Related work
In recent years, there have been a lot of studies from word embeddings to sentence embeddings and document embeddings. At the same time, from the existing research [27], sentence embeddings is very dependent on tasks. So we looked at existing technologies in two ways.

2.1 Sentence embeddings
Many studies have proposed various extensions of the Skip-gram model, which can capture specific semantic differences. Jamie et al. [6] proposed the Skip-thoughts model, which is composed of encoder-decoder based on RNN, to predict the sentences around a sentence. The limitation of this approach is that it requires a large amount of balanced corpus for training. The paper uses electronic books as training materials. Lajanugen et al. [7] improved the above model, proposed the quick-thoughts model, replaced the decoder with a classifier. Keshi et al. [8] built a semantic vector dictionary by introducing external knowledge base, and then use sentence embeddings for social media mining. Ruckle et al. [9] proposed the p-means method, a strong bag-of-word sentence embeddings baseline. Conneau et al. [10, 19, 24] proposed the InferSent model, trained the classifier on the SNLI data set, and used the bidirectional LSTM to complete the maximum pooling as the sentence encoder. But the effect of this supervised learning generalize to other areas is not good. Dhingra et al. [11] designed a model based on character combinations and learned vectorization from complex, non-normal language strings in the Twitter.

Most of the training materials used in the above work are long texts such as news, e-books. It is not suitable for the fragmented short texts like weibo.

2.2 Topic detection fusion of embeddings
The topic detection method is generally based on LDA. With the development of word embedding model in recent years, many researchers have done research on the combination of topic model and embeddings (especially sparse data). Moody [2] combined LDA and word2vec, and proposed the lda2vec model that can learn words, documents and subject vectors together. Zhao et al. [3] used the word embedding model to obtain global semantic information and built a hierarchical topic model. The topic structure was presented in the form of sub-topics. Zhao et al. [4] proposed a focused theme model based on information for the sparsity of short texts. Xun et al. [5] used the word-level association information contained in the word vector to model the subject correlation in the continuous word vector space. Liu et al. [6] learned the word vector and theme vector at the same time, and combined them into scene word vector, so that they could better represent the features of the same word in different scenes. Hashimoto et al. [25] used paragraph embeddings for topic detection in the public health field.

Existing technology do not perform well in sub-topics detection tasks, because there is not much distinction between sub-topics.

3. Method description

3.1 Overall framework
Figure 1. Sub-topics detection framework.

The process of sub-topics detection is shown in figure 1. First, use LDA for weibo data to get the topics. Then the word and the corresponding topic are trained together to get the word embeddings and the topic embeddings. By taking the cosine between the word embeddings and the topic embeddings as the weight value, the word embeddings of the target words under all topics are weighted and added, which is used to extend the topic information into the word embeddings and enhance the semantics of the word embeddings. The p-means method is used to merge the blog into the sentence embeddings, which is the characteristic value of the blog. Finally, the sub-topic clusters are obtained through k-means.

3.2 Word embeddings and topic embeddings

TWE model is an improved method based on Skip-gram model proposed by Liu et al. [5]. Learn the topic word embeddings (TWE) by using the topic model to assign potential topics to each word. The TWE model takes into account the polysemy of the word and is able to represent the characteristics of the same word under different topics.

This paper uses the TWE model to learn word embeddings and topic embeddings simultaneously. For each target word and its topic \(<w_i, t_i>\), TWE aims to maximize the average likelihood probability:

\[
L(\beta) = \frac{1}{M} \sum_{i} \sum_{k \in \mathbb{K}, c \neq 0} \log P_D(w_{j \in c} / w_i) + \log P_D(w_{j \in c} / t_i)
\]

Compared to Skip-gram, TWE also uses the topic of the target word \(t_i\). The basic approach is to treat each topic as a pseudo word that appears on all of the words assigned to the topic. So the topic embeddings will represent the overall semantics of the words under the topic.

3.3 Sentence embeddings

First, we combine the topic embedding with the word embedding. \(w^t\) is the vector representation of the word \(w\) under all topics. It can be expressed in the following form:
\[
\sum_{i \in T} w \cdot Pr \left( \| \cdot \| \right)
\]  

(2)

Cosine is used to measure the similarity between the target word and each topic, and is used as the weight. The resulting embeddings is used to represent the feature of the target word under all topics.

Then we combine the word vector into a sentence vector. Recently, a method to combine word embeddings into sentence embeddings has been proposed [9]. P-means is defined as:

\[
\left( \frac{x_1^p + ... + x_n^p}{n} \right)^\frac{1}{p} \quad p \in R \cup \{ \pm \infty \}
\]

(3)

It was the average operation when \( p=1 \), and was the maximum operation when \( p=+\infty \), the minimum operation when \( p=-\infty \). To effectively compress the overall information of the sentence, we use the above three operations (average, maximum and minimum) together.

Given a sentence that contains \( n \) words, each word embeddings has \( d \) dimensions:

\[
\|w\| = [w_1, \ldots, w_n] \in R^{n \times d}
\]

(4)

Different p-means are represented as \( H_p(W) \), so different p-means result connections can be represented as:

\[
s^{(i)} = H_{p_i} \left[ w^{(i)} \right] \oplus H_{p_k} \left[ w^{(i)} \right]
\]

(5)

Where, \( p_1, \ldots, p_k \) is \( K \) different p-means value. The resulting sentence vector is a vector of dimensions \( K \times d \).

4. Experiment and analysis

This paper mainly uses sina weibo data to verify the effect of the model. Validation is divided into two parts. First, test the effect of blog information represented by sentence embeddings on topic detection, and second, verify the accuracy of sub-topics detection under the same topic.

4.1 Experimental data set

In this paper, we test the validity of weibo information represented by sentence embeddings using weibo data from August to September 2014. Contains about 100,000 blogs. After preprocessing the weibo data, the topic model with 70 topics is obtained through LDA. Further learn the word embeddings and topic embeddings of the data set. Then generate the sentence embeddings for each blog. Later, take the sentence embeddings of blogs as features and use k-means to cluster. Set the dimension of word embeddings and topic embeddings to 300.

In addition, the effect of finding sub-topics was tested by using the Tianjin port explosion event. Contains 1,096 blogs. The data sets already manually label sub-topics, including accident descriptions, tribute to fire fighters, query handling, government response, and environmental issues. The theme model with topic number 5 is obtained through LDA. The word embeddings and topic embeddings dimensions are 50.

4.2 Evaluation criteria

For the ability of sentence embeddings to represent blog information, the evaluation criterion we use is artificial judgment. The method is based on the most topics assigned in each blog. Calculate the similarity with the sentence embeddings and each topic embedding, classify the blogs under the most similar topics, and then compare them with the benchmark.

For the detection effect of sub-topics, we use sub-topic labels of events manually marked, and evaluate the effect by calculating the accuracy of clustering results relative to manual labeling. The accuracy and the Rand Index (RI) were used as indicators to compare with other methods.

RI understands clustering as a series of decision-making processes. TP means to group similar documents into the same cluster, TN means to group dissimilar documents into different clusters, FP
means to group dissimilar documents into the same cluster, and FN means to group two similar documents into different clusters. RI actually calculated the ratio of the right decision. The calculation formula is:

\[
RI = \frac{TP + TN}{TP + TN + FN + FP}
\]  

(6)

4.3 Experimental results

In the experiment on the ability of sentence embeddings to represent blog information, the result of our proposed model and LDA was compared, and the consistency rate was 82%. Partial results of inconsistencies are shown in table 1. The results show that the proposed model is more semantic.

Table 1. Evaluation results compared with LDA.

| Blog | LDA         | Our model      |
|------|-------------|----------------|
| Obstetrician in Shandong TV hit, in order to better the doctor white angels of positive energy to screen, from the second hospital of Shandong university yong-ping xu, director of the maternity and the obstetrician Liu Yuan guest of Qilu hospital of Shandong university Qilu nets, as an obstetrician heat, with net friend to share the real life of the obstetrician. | Hospital Net friend Found Doctor Event | Show Live Tonight Doctor Video |
| [Shijiazhuang has reached 60 days above grade two this year] As of August 14, the provincial capital had received five first-grade days and 55 second-grade days, with an increase of 15.9% in the percentage of excellent days compared with the same period last year, and a decrease in the number of days with moderate or severe pollution. | Time Internet Surf Using Suitable for Waste | Air Quality Instant noodles A series of Purifier |
| I think one day, I would like to cut my short hair cleanly, wear a big T-shirt, light blue jeans, carry a big travel bag, across China. do not affect the card. | Wear Clothes Like Match clothes Good look | Load On the load Grow up Life Future |

Then, we taking the Tianjin port explosion event as an example to show the effect of the model on the detection of sub-topics. Table 2 shows partial results of the detection of sub-topics of the event. Table 3 shows the comparison between our proposed model and other algorithms, in which our proposed model has an accuracy rate of 85%.

Table 2. Detection result of Tianjin port explosion event.

| Blog | Subtopic label | Clustering results |
|------|---------------|--------------------|
| The fire department of the ministry of public security said that nine squadrons of the tianjin fire brigade and three full-time teams from the port bureau of the port authority rushed to the scene to fight a fire that broke out in the piling of dangerous chemicals by the ruihai logistics of tianjin | accident descriptions | consistent |
binhai new area port group. An explosion occurred at about 23:30.

At 11:00 PM on the 12th, tianjin port ruihai company owned dangerous goods warehouse explosion. By 10:00 PM on the 13th, nine firefighters had been killed and many others were missing. Regardless of the fire or disaster areas, where there is danger, where there is a fearless firefighter figure. Today, we are sending out a tweet to all the heroes of the fire service!

It was cruel to send these children in firefighter uniforms to the fire without any knowledge of using drones or remote sensing to assess the risk. It doesn't matter how many years a firefighter should be professional, not militarized. It's human life.

| Algorithms          | RI   | Accuracy | F1   |
|---------------------|------|----------|------|
| k-means             | 0.428| 0.481    | 0.478|
| LDA                 | 0.792| 0.703    | 0.701|
| BTM                 | 0.784| 0.805    | 0.797|
| Word embeddings     | 0.842| 0.813    | 0.804|
| Our model           | 0.861| 0.854    | 0.839|

5. Conclusion
It is of great practical significance to monitor the public opinion events of major events and natural disasters and analyze their trends and dissemination.

Sub-topics refer to different aspects within the topic, and detecting sub-topics of the event can help us to get the whole picture of the event. As a special task of topic detection, sub-topic detection is featured by the fact that all the documents processed are reports or descriptions of the same event, which leads to high similarity between documents. How to ensure that the sub-topics detected are sufficiently different is a big problem. While vectorization can preserve the semantics of the text to a great extent, it can also improve the comprehensibility of sub-topics.

Based on the demand and research status of public opinion event analysis, this paper proposes a new sub-topic discovery model, which uses sentence embeddings as the feature of blog post to get the cluster of sub-topics. The experimental data of weibo can prove that this model performs well in the task of finding sub-topics. Compared with previous methods, the accuracy is improved.

In the next step, we will build on this to incorporate other approaches and consider using other topic models that are more suitable for the short text. It can also make use of the characteristics of weibo to introduce factors such as time factors to generate sentence embeddings with more information on the basis of accuracy and comprehensibility.
Acknowledgement
This research was financially supported by National Key Fundamental Research (NO. 2017YFB0803303) and National Natural Science Foundation of China (NO. 61732022, 61732044).

References
[1] Lin W Q, Deng L, Ding Z Y, et al. Hierarchical Dynamic Community Detection by Parallel Computing. Chinese Journal of Computers, 35(8):1712, 2012.
[2] Moody C E. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. 2016.
[3] Zhao H, Du L, Buntine W, Zhou M Y. Inter and Intra Topic Structure Learning with Word Embeddings. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:5892-5901, 2018.
[4] Zhao H, Du L, Buntine W. A word embeddings. informed focused topic model. In ACML, pp. 423–438, 2017.
[5] Xun G, Li Y, Zhao W, Gao J, Zhang A. A correlated topic model using word embeddings. In IJCAI, pp. 4207–4213, 2017.
[6] Liu Y, Liu Z, Chua T S, et al. Topical word embeddings. Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015:2418-2424.
[7] Logeswaran L, Lee H. An efficient framework for learning sentence representations. 2018.
[8] Keshi I, Yu S, Yoshino K, et al. Semantically readable distributed representation learning for social media mining[C]/ International Conference on Web Intelligence. ACM, 2017:716-722.
[9] Rücklé A, Eger S, Peyrard M, et al. Concatenated p-mean Word Embeddings as Universal Cross-Lingual Sentence Representations. 2018.
[10] Conneau A, Kiela D, Schwenk H, et al. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. 2017:670-680.
[11] Dhingra B, Zhou Z, Fitzpatrick D, et al. Tweet2Vec: Character-Based Distributed Representations for Social Media. 2016.
[12] Wang S, Zhang J, Zong C. Empirical Exploring Word-Character Relationship for Chinese Sentence Representation[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2018, 17(3):1-18.
[13] Cer D, Yang Y, Kong S Y, et al. Universal Sentence Encoder. 2018.
[14] Subramanian S, Trischler A, Bengio Y, et al. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. 2018.
[15] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. 2018.
[16] Lin Z, Feng M, Santos C N D, et al. A Structured Self-attentive Sentence Embedding. 2017.
[17] Mrkšić N, Vulić I, Séaghdha D O, et al. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. 2017.
[18] Qiang J, Chen P, Wang T, et al. Topic Modeling over Short Texts by Incorporating Word Embeddings. 2016:363-374.
[19] Conneau A, Kruszewski G, Lample G, et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. 2018.
[20] Shen D, Wang G, Wang W, et al. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. 2018.
[21] Liu P, Qiu X, Huang X. Learning context-sensitive word embeddings with neural tensor skip-gram model. International Conference on Artificial Intelligence. AAAI Press, 2015:1284-1290.
[22] Dai Q N, Nguyen D Q, Modi A, et al. A Mixture Model for Learning Multi-Sense Word Embeddings. Joint Conference on Lexical and Computational Semantics. 2017.
[23] Ma S, Sun X, Li W, et al. Word Embedding Attention Network: Generating Words by Querying Distributed Word Representations for Paraphrase Generation. 2018.
[24] Conneau A, Kruszewski G, Lample G, et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties[J]. 2018.

[25] Hashimoto K, Kontonatsios G, Miwa M, et al. Topic Detection Using Paragraph Vectors to Support Active Learning in Systematic Reviews[J]. Journal of Biomedical Informatics, 2016, 62(C):59-65.

[26] Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings. 2016.

[27] Iyyer M, Boyd-Graber J, Claudino L, et al. A Neural Network for Factoid Question Answering over Paragraphs[C]/ Conference on Empirical Methods in Natural Language Processing. 2014:633-644.