Neurocomputing 174 (2016) 806–814

Contents lists available at ScienceDirect

Neurocomputing

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

Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification

Peng Wang [a,*], Bo Xu [a], Jiaming Xu [a], Guanhua Tian [a], Cheng-Lin Liu [a,b], Hongwei Hao [a]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China
[b] National Laboratory of Pattern Recognition (NLPR), Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 4 May 2015
Received in revised form
22 June 2015
Accepted 30 September 2015
Communicated by Jinhui Tang
Available online 9 October 2015

Keywords:
Short text
Classification
Clustering
Convolutional neural network
Semantic units
Word embeddings

ABSTRACT

Text classification can help users to effectively handle and exploit useful information hidden in large-scale documents. However, the sparsity of data and the semantic sensitivity to context often hinder the classification performance of short texts. In order to overcome the weakness, we propose a unified framework to expand short texts based on word embedding clustering and convolutional neural network (CNN). Empirically, the semantically related words are usually close to each other in embedding spaces. Thus, we first discover semantic cliques via fast clustering. Then, by using additive composition over word embeddings from context with variable window width, the representations of multi-scale semantic units[1] in short texts are computed. In embedding spaces, the restricted nearest word embeddings (NWEs)[2] of the semantic units are chosen to constitute expanded matrices, where the semantic cliques are used as supervision information. Finally, for a short text, the projected matrix[3] and expanded matrices are combined and fed into CNN in parallel. Experimental results on two open benchmarks validate the effectiveness of the proposed method.

1. Introduction

The classification of short texts, such as search snippets, micro-blogs, product reviews, and short messages, plays important roles in user intent understanding, question answering and intelligent information retrieval [1]. Since short texts do not provide enough contextual information, the data sparsity problem is easily encountered [2]. Thus, the general methods based on bag-of-words (BoW) model cannot be directly applied to short texts [1], because the BoW model ignores the order and semantic relations between words. How to acquire effective representations of short texts to enhance the categorization performance has been an active research issue [2,3].

Conventional text classification methods often expand short texts using latent semantics, learned by latent Dirichlet allocation (LDA) [4] and its extensions. Phan et al. [3] presented a general framework to expand short and sparse texts by appending topic names, discovered using LDA over Wikipedia. Sahami and Heilman [5] enriched text representation by web search results using the short text segment as a query. Furthermore, Yan et al. [6] presented a variant of LDA, dubbed biterm topic model (BTM), especially for short text modeling to alleviate the data sparsity problem. However, these methods still consider a text as BoW. Therefore, they are not effective in capturing fine-grained semantics for short texts modeling.

More recently, deep learning based methods have drawn much attentions in the field of natural language processing (NLP), which mainly evolved into two branches. One is to learn word embeddings by training language models [7–10], and another is to perform semantic composition to obtain phrase or sentence level representation [11,12]. Word embeddings, also known as distributed representations, typically represent words with dense, low-dimensional and real-valued vectors. Each dimension of the vectors encodes a different aspect of words. In embedding spaces, semantically close words are likely to cluster together and form semantic cliques. Moreover, the embedding spaces exhibit linear structure that the word embeddings can be meaningfully combined using simple vector addition [9].

In this paper, we aim to obtain the semantic representations of short texts and overcome the weakness of conventional methods. Similar to Li et al. [13] that cluster indicators learned by non-negative spectral clustering are used to provide label information for structural learning, we develop a novel method to model short texts using word embeddings clustering and convolutional neural network (CNN). For concision, we abbreviate our methods to

* Corresponding author.

[1] Semantic units are defined as n-grams which have dominant meaning of text. With n varying, multi-scale contextual information can be exploited.

[2] In order to prevent outliers, a Euclidean distance threshold is preset between semantic cliques and semantic units, which is used as restricted condition.

[3] The projected matrix is obtained by table looking up, which encodes Unigram level features.

http://dx.doi.org/10.1016/j.neucom.2015.09.096
0925-2312/

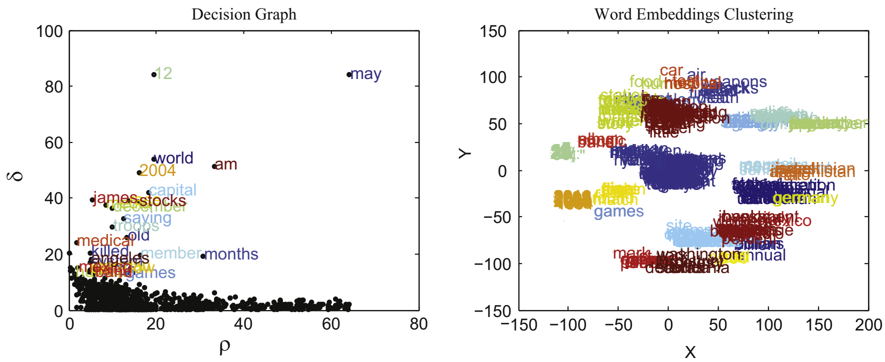

Fig. 1. Fast clustering based on density peaks of embeddings.

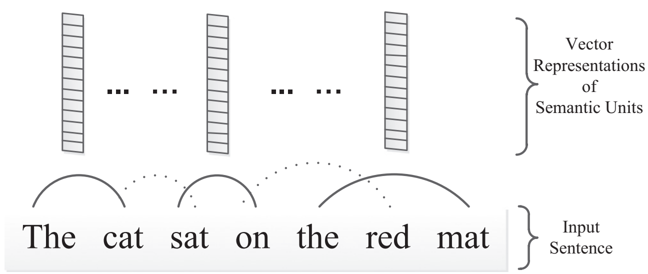

Fig. 2. The detection of semantic units.

CCNN, as Clustering and CNN are employed. Particularly, the fast clustering algorithm, based on density peaks searching [14], is first utilized to discover the semantic cliques in embedding spaces, as shown in Fig. 1. Then, the component-wise additive composition is performed over word embeddings, from context with variable length, to compute the representations of semantic units appearing in short texts, as shown in Fig. 2. The semantic units are used to calculate Euclidean distance with each semantic clique, and their nearest word embeddings (NWEs) can be found. In our framework, the NWEs that meet the preset threshold of Euclidean distance are chosen to constitute the expanded matrices for short texts enrichment, otherwise simply dropout. In this stage, the semantic cliques are used as supervision information to detect precise semantics. Finally, a CNN with one convolutional layer followed by a K-max pooling layer is trained under the cross entropy objective, which is optimized with mini-batches of samples iteratively by back propagation (BP).

The motivation of the proposed method is to introduce semantic knowledge and expand short texts by related word embeddings, which is pre-trained over large-scale external corpus. To preserve the semantics in original short texts, we integrate text understanding and vectorization into a joint framework. As shown in Fig. 2, for the input short text "*The cat sat on the red mat*", three semantic units can be detected with different window width. These multi-scale semantic information is leveraged to expand the short text, and its context is fully exploited.

The main contributions of this paper are summarized as follows:

(1) The density peaks searching based clustering method is utilized to discover semantic cliques, which are used as supervision information to extract fine-tuned semantics.
(2) Multi-scale semantic units are defined and their representations are calculated by using a one-dimensional convolution-like operation.
(3) In embedding spaces, the restricted NWEs of semantic units are discovered to produce expanded matrices. Then, the projected matrix and the expanded matrices are simply combined and fed into a CNN to extract high-level features.

Experiments are conducted on Google snippets [3] and TREC [15] to validate the effectiveness of our method.

The rest of this paper is organized as follows. Section 2 gives a brief review of related works. Section 3 introduces the theoretical foundation of our work, including semantic composition and word embeddings clustering. Section 4 defines the relevant operators and hierarchies of the framework. Section 5 presents our experimental results. Finally, concluding remarks are offered in Section 6.

2. Related works

In order to overcome the data sparsity problem in short texts representations, many popular solutions have been proposed. Based on external Wikipedia corpus, Phan et al. [3] proposed a method to discover hidden topics using LDA and expand short texts. Zhou et al. [16] exploited semantic information from Wikipedia to enhance the question similarity in concept space. Chen et al. [2] proved that leveraging topics at multiple granularity can model short texts more precisely.

In recent years, neural networks (NNs) relevant methods have been used to model languages with promising results, and word embeddings can be learned meanwhile [17]. Mikolov et al. [9] introduced the continuous Skip-gram model that is an efficient method for learning high quality word embeddings from large-scale unstructured text data. Furthermore, various pre-trained word embeddings are publicly available, and many composition-based methods are proposed to induce semantic representations of texts.

To obtain sentence-level representations of texts, NNs related works can be divided into two types, which are respectively used for universal tasks and special tasks. For the former, Le and Mikolov [12] proposed the paragraph vector to learn a fixed-size feature representation for documents with variable length. Kalchbrenner et al. [18] introduced the dynamic convolutional neural network (DCNN) for modeling sentences, which is the most related work to our study. In that work, dynamic k-max pooling is utilized to capture global features without relying on parse tree. Based on convolutional architecture, Kim [19] proposed a simple improvement that two input channels are used which allow the employment of dynamic-updated and static word embeddings simultaneously. These methods can be used to generate semantic representations of texts for various tasks.

For the latter, Zeng et al. [20] developed a deep convolutional neural network (DNN) to extract lexical and sentence level features, which are used for relation classification. Socher et al. [21] proposed the recursive neural network (RNN) that has proven to be effective in sentiment prediction. In order to reduce the overfitting problem of neural network, especially trained on small data set, Hinton et al. [22] used random dropout to prevent complex co-adaptations.