Knowledge-Based Systems 106 (2016) 220–230

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Learning distributed word representation with multi-contextual mixed embedding

Jianqiang Li, Jing Li, Xianghua Fu*, M.A. Masud, Joshua Zhexue Huang

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 22 November 2015
Revised 22 May 2016
Accepted 23 May 2016
Available online 24 May 2016

Keywords:
Word embedding
Distributed word representation
Word2vec
Natural language processing

ABSTRACT

Learning distributed word representations has been a popular method for various natural language processing applications such as word analogy and similarity, document classification and sentiment analysis. However, most existing word embedding models only exploit a shallow slide window as the context to predict the target word. Because the semantic of each word is also influenced by its global context, as the distributional models usually induced the word representations from the global co-occurrence matrix, the window-based models are insufficient to capture semantic knowledge. In this paper, we propose a novel hybrid model called mixed word embedding (MWE) based on the well-known word2vec toolbox. Specifically, the proposed MWE model combines the two variants of word2vec, i.e., SKIP-GRAM and CBOW, in a seamless way via sharing a common encoding structure, which is able to capture the syntax information of words more accurately. Furthermore, it incorporates a global text vector into the CBOW variant so as to capture more semantic information. Our MWE preserves the same time complexity as the SKIP-GRAM. To evaluate our MWE model efficiently and adaptively, we study our model on linguistic and application perspectives with both English and Chinese dataset. For linguistics, we conduct empirical studies on word analogies and similarities. The learned latent representations on both document classification and sentiment analysis are considered for application point of view of this work. The experimental results show that our MWE model is very competitive in all tasks as compared with the state-of-the-art word embedding models such as CBOW, SKIP-GRAM, and GloVe.

1. Introduction

A word representation is a mathematical objects associated with each word. Learning word representations is an essential task of the natural language processing (NLP). The area of Natural Language Processing (NLP) has evolutionally three different representation paradigms, i.e., the bag-of-words, bag-of-concepts, and bag-of-narratives models [1–3]. Traditionally, the default word representation approach regards words as a one-hot vector, which shares the same size of the vocabulary, and most of which locations are 0 and only one dimension is 1. This storage structure makes the words suffer from great data sparseness and high dimensionality. Moreover, it is difficult to capture the complex linguistic characteristics of words in the large corpus. To address this issue, a lot of semantic-based approaches are proposed in literature, which can be broadly grouped into three categories[3]: techniques that leverage on external knowledge, e.g., ontologies (taxonomic NLP) [4–7] or semantic knowledge bases (noetic NLP) [8, 9], and methods that exploit only intrinsic semantics of documents (endogenous NLP) [5,10–13]. Nowadays, there are two families of the word representation are widely adopted: 1) distributional representation, and 2) distributed representation [14]. Distributional word representations follow the distributional hypothesis of Harris [15]: “linguistic items with similar distributions have similar meanings”. Many distributional word representation methods have been explored in NLP community in the past twenty years. On one end of the spectrum, words are grouped into clusters based on their contexts [16,17]. On the other end, words are represented as a very high dimensional and sparse vectors in which each entry is a measure of the association between the word and a particular context [18,19], SVD [63] or LDA [64] techniques are used to reduce the dimensionality of the sparse vector space. Distributional word representations are typically based on the textual contexts in which it has been observed, and usually induced from a global co-occurrence matrix of the words and their context, such as the term–document matrix, the word–context matrix, and the pair–pattern matrices matrix [19].

* Corresponding author.
E-mail addresses: lijq@szu.edu.cn (J. Li), muziqingqing@yahoo.com (J. Li), fuxh@szu.edu.cn (X. Fu), masud@szu.edu.cn (M.A. Masud), zx.huang@szu.edu.cn (J.Z. Huang).

http://dx.doi.org/10.1016/j.knosys.2016.05.045
0950-7051/© 2016 Elsevier B.V. All rights reserved.

Most recently, distributed word representations draw more attention for better performance in a broad range of natural language processing tasks, ranging from speech tagging [20], named entity recognition [14], part-of-speech tagging, parsing [21], and semantic role labeling [22], phrase recognition [21], sentiment analysis [23], paraphrase detection [24], to machine translation [25]. Different from the distributional word representation, distributed word representation methods are proposed to learn a dense, low-dimensional, and real value vector space representation of words by analyzing their usage in large corpora of textual documents [26–29]. These representations also are called as word embeddings or word vectors. Each dimension of the embedding represents a latent feature of the word, hopefully capturing useful syntactic and semantic properties.

Although distributed representations for words have become successful paradigms, especially for statistical language modeling [27, 29–31], most of them use neural network architectures to learn word embeddings, and the weaknesses are also obvious, such as the extremely long training times [26], and the window-based inducing process with only local contexts about 2–10 words [22,14]. There are enormous opportunities to introduce new methods for improving the expensive time-consuming issue on large corpora, such as structured the vocabulary into a tree with words at the leaves which allows exponentially faster computation of word probabilities and their gradients [29, 31], trained the language model with noise-contrastive estimation [32], and implemented the language models on a large distributed framework [33]. Most recently, Mikolov designed a new tool, i.e., word2vec, to learn distributed word representation very quickly, which can take less than a day to learn high quality word vectors from a 1.6 billion words dataset [34, 35]. However, there are few works about the drawback of the window-based inducing process of word embedding approaches. A pioneering work on word embedding learning with multiple level contexts is Huang's neural network architecture [36], where the word embeddings are learned by incorporating both local context and global document context, and the neural network architecture is based on the recursive neural network. Another work is GloVe [37], which is a global log-bilinear regression model that combines the advantages of the global matrix factorization methods and the local context window methods.

In this paper, we propose a mixed word embedding (MWE) model to improve the word embedding learning of the word2vec. As we known, the word2vec [34] includes two model architectures, one is the continuous bag-of-words model (CBOW), and the other is the continuous skip-gram model (SKIP-GRAM). CBOW uses a word's context words in a surrounding window to predict the word, while SKIP-GRAM uses a word to predict its surrounding words. We can see that SKIP-GRAM and CBOW are the opposite of each other, and they can improve each other. So we integrate SKIP-GRAM and CBOW together in our MWE model, and add a global text vector into MWE. Furthermore, we develop an incremental window algorithm of our MWE model. At each iteration of the word embedding learning, our algorithm predicts the word vector of each word in the local context according a word's vector based on SKIP-GRAM at first, and then the new word vectors of the local context and the global text vector are used to predict the word's vector again with an incremental window. We compare our MWE with state-of-the-art distributed word representation methods, such as SKIP-GRAM, CBOW [34], and GloVe [37] in several NLP tasks, such as word similarity, word analogy, document classification, and sentiment analysis, the experimental results show that our MWE can get the best performance in the most of tasks. To the best of our knowledge, MWE is the first mixed embedding model which integrated SKIP-GRAM and CBOW, and combine the local context and the global context. The main contributions of this paper are four-folds:

(1) We combine the SKIP-GRAM model with the CBOW architecture for words training in each local context window. Since the SKIP-GRAM and the CBOW are the opposite of each other, and they can improve each other, it is a good way to combine them to get high quality word vectors.
(2) We assigned a global text vector for each document, and concatenated it to the CBOW model for words training. The text vector is computed by the weighted average of all word vectors in the text, and it will be updated with word vectors training in each context window. After training, word embeddings and text vectors will be simultaneously got by the MWE model, they can also be put into many other deep learning structures to finish complex NLP tasks.
(3) We propose an incremental widows learning algorithm of our MWE model. The standard CBOW uses a sum pooling layer of all words in the local context window to speed up its training process. The size of the local context window is fixed, and the words' order of the local context is ignored. Our incremental learning algorithm predicts each word of the local context in order, and predicts the word with a variable and increment window. With the increment learning mode, our algorithm considers not only the words' order of the local context, but also preserves the same time complexity as the SKIP-GRAM.
(4) Most of models that evaluate word representation only on one perspective, either word linguistic analysis or some supervised learning tasks such as word segmentation, text classification and so on. We evaluate our model on linguistic and application perspectives with both English and Chinese dataset. The experimental results show that our MWE model is very competitive in all tasks as compared with the state-of-the-art word embedding models such as CBOW, SKIP-GRAM, and GloVe.

The remainder of this paper proceeds as follows: In Section 2 , we describe some previous works related to our study, including developments of the word representation and neural network language model, word embedding evaluations, and our method. Section 3 illustrates architecture of the MWE model as well as its algorithm and detail implementation. Experiments and assessments are reported in Section 4 . Finally, we conclude the paper and introduce some future works.

2. Related works

Unsupervised word representations have been used in previous NLP works, and have demonstrated improvements in generalization accuracy on a variety of tasks. Conventionally, supervised lexicalized NLP approaches take a word and convert it to a symbolic ID, which is then transformed into a feature vector using a one-hot representation. However, the one-hot representation of a word suffers from data sparsity. Moreover, at test time, the model cannot handle words that do not appear in the labeled training data. These limitations of one-hot word representations have prompted researchers to investigate unsupervised methods for inducing word representations over large unlabeled corpora. One common approach to inducing unsupervised word representation is to use clustering. This technique was used by a variety of researchers [17,65–68]. This leads to a one-hot representation over a smaller vocabulary size. Neural language models [26,38,27, 39], on the other hand, induce dense real valued low dimensional word embeddings using unsupervised approaches.

Generally, researchers choose neural network to train language model and generate useful word representation for a wide range of NLP tasks [14,27,38,40]. The pioneering work on word embeddings