Information Sciences 430–431 (2018) 216–227

Contents lists available at ScienceDirect

Information Sciences

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

Entity mention aware document representation

Hongliang Dai, Siliang Tang*, Fei Wu, Yueting Zhuang

Zhejiang University, Hangzhou, Zhejiang, China

ARTICLE INFO

Article history:
Received 19 February 2017
Revised 10 October 2017
Accepted 17 November 2017
Available online 20 November 2017

Keywords:
Distributed representation
Text clustering
Text classification
Entity linking

ABSTRACT

Representing variable length texts (e.g., sentences, documents) with low-dimensional continuous vectors has been a topic of recent interest due to its successful applications in various NLP tasks. During the learning process, most of existing methods tend to treat all the words equally regardless of their possibly different intrinsic nature. We believe that for some types of documents (e.g., news articles), entity mentions are more informative than ordinary words and it can be beneficial for certain tasks if they are properly utilized. In this paper, we propose a novel approach for learning low-dimensional vector representations of documents. The learned representations captures information of not only the words in documents, but also the entity mentions in documents and the connections between different entities. Experimental results demonstrate that our approach is able to significantly improve text clustering, text classification performance and outperform previous studies on the TAC-KBP entity linking benchmark.

1. Introduction

Learning low-dimensional vector representations for documents can be an effective approach in many NLP tasks and has the potential to outperform traditional representation methods like bag-of-words (BoW) [9,15,19]. However, despite the fact that different words have different properties and are of different importance to the whole document, such methods always tend to treat all words equally.

In particular, existing document representation learning methods do not distinguish named entity mentions (e.g., the mention of “Hillary Clinton” in a document) with ordinary words. Entity mentions occur frequently in various types of documents and are usually more informative than most ordinary words. In news articles, mentioned entities such as persons, organizations and locations provide essential information about *who* and *where* of the reported events. Two documents are more likely to have similar content if some of the entities mentioned by them are the same. Sometimes, knowing what entities are mentioned, we can even infer the possible topics of a document. For example, if a news article mentioned both “Hillary Clinton” and “Donald Trump”, then we know there is a high probability that this article is about the US presidential election of 2016. Thus entity mention information can help to better capture the semantic similarities between documents while learning document representations. Moreover, different entities may be related with each other. For example, a person has connections with many other persons and is related to many locations and organizations. Documents that mention different but related entities are also more likely to have similar content. In order to leverage this property, the relatedness between different entities should also to be considered while learning representations for documents.

* Corresponding author.
E-mail addresses: hldai@zju.edu.cn (H. Dai), siliang@zju.edu.cn (S. Tang), wufei@zju.edu.cn (F. Wu), yzhuang@zju.edu.cn (Y. Zhuang).

https://doi.org/10.1016/j.ins.2017.11.032
0020-0255/

Therefore, we believe that it is possible to improve the quality of document representations by capturing the entity mention information of documents and the relatedness between different entities. Document representations learned with such information may achieve better performance when applied to tasks such as text clustering and text classification. They are also very suitable for the task of entity linking, which aims to map the mentions in a document to their referred entities in the referent knowledge base, since existing research [6,10,21,30] has already shown that while performing entity linking for a mention, other mentions in the same context can be particularly helpful for the inference.

In this paper, we propose a novel approach to learn distributed representations of documents that are aware of the entity mentions in documents. We name our approach EMADR (Entity Mention Aware Document Representations). EMADR generalizes the PV-DBOW model proposed by Le and Mikolov [19] to make it possible to incorporate multiple types of related information into document representations. The learned document representations captures three types of information: what words are used in each document, what entities are mentioned in each document and the relatedness between different entities.

The main contributions of this paper are:

- We propose EMADR, which to the best of our knowledge, is the first document representation learning method that leverages entity mention information.
- We apply EMADR to entity linking with a neural network model. Compare with some existing neural network methods [12,33] for this task, it has the advantage of only requiring a small amount of training data.
- We study the performance of EMADR by conducting experiments on text clustering, text classification and entity linking. We find that EMADR is able to significantly improve text clustering and classification performance. Its application in entity linking also beats previous studies on the TAC-KBP entity linking benchmark.

The rest of this paper is structured as follows: In Section 2 we discuss the technical details of learning document representations with our approach. Section 3 shows how we apply the learned representations to the task of entity linking. In Section 4, we conduct a series of experiments on text clustering, text classification and entity linking. Finally, we introduce some related works in Section 5 and dummyTXdummy- concludes our work in Section 6.

2. Entity mention aware document representations

As previously mentioned in the introduction, for each document, we want its representation to capture both what words are used and what entities are mentioned. We also aim to capture the relatedness between different entities so that the representations of documents with different but related entities may also be similar. In order to do this, we generalize the well-known document representation learning model PV-DBOW [27] by introducing the concept of *prediction lists*. Then we employ this idea to learn document representations based on constructing three prediction lists.

2.1. Embedding method based on prediction lists

We start by introducing the PV-DBOW model. PV-DBOW represents each document with a dense vector that is trained to model the distribution of words in the document. Given a set of documents D and a set of words W, It uses a softmax to model the probability of observing word w in document d:

$$p(w|d) = \frac{\exp(v'^{\top}_w v_d)}{\sum_{\hat{w}\in W} \exp(v'^{\top}_{\hat{w}} v_d)}, \tag{1}$$

where v_d is the vector representation of d, v'_w is the weight vector with respect to w. Eq. (1) has the property that if a word w occurs frequently in document d, then $v'^{\top}_w v_d$ should be large, which usually means v_d will be similar with v'_w.

Let $w^d_1, w^d_2, \ldots, w^d_{N(d)}$ be the sequence of words in document $d \in D$, where $N(d)$ is the number of words in d. The objective of PV-DBOW is to maximize the log probability

$$I = \sum_{d\in D} \sum_{i=1}^{N(d)} \log p(w^d_i|d). \tag{2}$$

The document representations learned with Eqs. (1) and (2) will preserve second-order proximity [34], which means that if two document use similar words, then their corresponding representations will also be similar.

In order to generalize this model, we use $\langle x, y\rangle$ to denote a positive sample of observing y given x, which means $p(\langle x, y\rangle) = p(y|x)$. For each document $d \in D$, we add $\langle d, w^d_1\rangle, \langle d, w^d_2\rangle, \ldots, \langle d, w^d_{N(d)}\rangle$ in a list L, then maximizing Eq. (2) equals to maximizing the log probability of observing all the samples in L. We call $\langle x, y\rangle$ a prediction sample and L a prediction list. This shows that we can get the same objective as the PV-DBOW model based on a prediction list constructed from the documents.

Moreover, we can also train the model based on the constructed list. Suppose we randomly draw T samples from L and denote them as $\langle \hat{d}_1, \hat{w}_1\rangle, \langle \hat{d}_2, \hat{w}_2\rangle, \ldots, \langle \hat{d}_T, \hat{w}_T\rangle$. Then it can be easily shown that when T is sufficiently large,

$$\sum_{i=1}^{T} \log p(\langle \hat{d}_i, \hat{w}_i\rangle) \approx \frac{T}{\sum_d N(d)} \cdot \sum_{d\in D} \sum_{i=1}^{N(d)} \log p(w^d_i|d). \tag{3}$$