Knowledge-Based Systems 102 (2016) 76–86

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Improving short text classification by learning vector representations of both words and hidden topics

Heng Zhang[a,*], Guoqiang Zhong[b]

[a] Interactive Digital Media Technology Research Center, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, P.R. China

[b] Department of Computer Science and Technology, Ocean University of China, 238 Songling Road, Qingdao 266100, P.R. China

ARTICLE INFO

Article history:
Received 13 September 2015
Revised 25 March 2016
Accepted 27 March 2016
Available online 30 March 2016

Keywords:
Short texts
Topic model
Data enrich
Word and topic vectors

ABSTRACT

This paper presents a general framework for short text classification by learning vector representations of both words and hidden topics together. We refer to a large-scale external data collection named "corpus" which is topic consistent with short texts to be classified and then use the corpus to build topic model with Latent Dirichlet Allocation (LDA). For all the texts of the corpus and short texts, topics of words are viewed as new words and integrated into texts for data enriching. On the enriched corpus, we can learn vector representations of both words and topics. In this way, feature representations of short texts can be performed based on vectors of both words and topics for training and classification. On an open short text classification data set, learning vectors of both words and topics can significantly help reduce the classification error comparing with learning only word vectors. We also compared the proposed classification method with various baselines and experimental results justified the effectiveness of our word/topic vector representations.

1. Introduction

The popular use of the Internet demands the technology for short text classification [1–3] to deal with the daily/history big data such as search snippets, communication messages, product titles and so on. The bag-of-words (BOW) feature representation has achieved satisfactory results for the analysis of normal text/document based on machine learning methods [4,5] such as SVM, kNN, maximum entropy and so on. But the BOW feature has very little sense about semantics of words and so fails to achieve desired classification accuracy on short texts [6] which do not provide sufficient word co-occurrence or context shared information for effective similarity measure. Therefore, it is necessary to conduct in-depth study on feature representations for short texts.

To tackle the data sparseness problem of short texts, various methods have been proposed in literatures. Zelikovitz et al. [7] described a method for improving the classification of short text strings using a combination of labeled training data plus a secondary corpus of unlabeled but related longer documents. Hu et al. [8] enriched document representations with Wikipedia concepts and category information by mapping text documents to Wikipedia concepts, and further to Wikipedia categories. For query classification, Cao et al. [9] used neighboring queries and their corresponding clicked URLs (Web pages) in search sessions as the context information and incorporated the context information into the problem of query classification by using conditional random field (CRF) models. He et al. [10] employed a supervised-learning method to learn hint verbs, and considered URL and title information to classify snippets into three coarse categories. Dhillon et al. [11] studied a certain spherical k-means algorithm for clustering large sparse text data. Phan et al. [12] derived a set of hidden topics through topic model LDA from one large existing Web corpus for short text expansion. Vo et al. [13] also used LDA model for topic analysis but presented new methods for enhancing features by combining external texts modeled from various types of universal datasets.

Recently, word vector representations have been demonstrated to be able to produce outstanding results in some short text classification work such as sentiment analysis [14–17]. Word vectors can be learned via language modeling [18] or encoding word meaning (semantics) [14] with a probabilistic modeling approach. Based on word vectors, the text feature can be represented as the average of vector representations(*mean representation vector*) for all the words in the document [14] or learned in an unsupervised framework based on continuous distributed vector representations for the document [17]. The text feature can be used for many document analysis work such as clustering, classification and retrieval.

* Corresponding author. Tel.: +8615210237582.
E-mail addresses: heng.zhang@ia.ac.cn (H. Zhang), gqzhong@ouc.edu.cn (G. Zhong).

http://dx.doi.org/10.1016/j.knosys.2016.03.027
0950-7051/

Besides of document analysis, word vectors can be also used in NLP applications such as named entity recognition, word sense disambiguation, parsing, tagging and machine translation.

Following the trend of short text enriching and word vector learning, we attempt to learn vector representations of both words and topics to improve short text classification. During the learning of topic model, each word from texts on the corpus is assigned with a topic [12] and these word-topic assignments (each pair of the word and its assigned topic is denoted as a word-topic assignment) at the end of topic learning are used to enrich texts on the corpus (corpus with text enriching is denoted as enriched corpus). Then both topic and word vectors can be learned on the enriched corpus by modifying traditional methods. In short text classification, short texts are enriched in a similar way as the text enriching on corpus: by doing topic inference on short texts, each word of short texts is also assigned with a topic [12] and these word-topic assignments at the end of topic inference are used to enrich short texts (short texts enriched with topics are denoted as enriched short texts). Then, features of short texts can be represented based on vectors of not only words but also topics. By exploiting benefits of topic models for text enriching and the vector based feature representation, our method can achieve the superior performance over many exiting references such as topic feature, knowledge enriching, and traditional word vector learning. The idea of learning word vectors together with topics is similar to models of topical word embeddings (TWE) [42]. TWE models view topics as pseudo words to predict contextual words, while we view topics as new words in the text and learn vectors of words and topics more interactively. Enriching short texts with topics has been successfully proposed in a different way [12] with us. In our framework, enriching texts with topics can not only overcome the data sparseness problem of short texts but also make it possible to learn vector representations of words and topics together.

The rest of this paper is organized as follows: in Section 2, we give a brief review of related works on short text classification. Section 3 gives an overview of the proposed theoretical framework. Section 4 describes the topic modeling with LDA and text enriching with topics. Section 5 gives the algorithm on learning vector representations of both words and topics. Section 6 validate the proposed system over an open data set and Section 7 offers concluding remarks and future work.

2. Related works

In this section, we briefly summarize related works on the basic text representation i.e. the BOW model, the content/feature extension for short texts, word-vector based text classification methods which are more related to ours.

2.1. BOW text representation

The use of popular text representation known as BOW can trace back to Harris's 1954 article on "Distributional structure" [23] and widely used for text classification, clustering and retrieval. In the BOW representation, it is common to weigh terms by various schemes such as TF, TF-IDF [24] and its variants [25–27]. Using these word-based feature representations in the high dimensional space, a great many text classification techniques have been proposed such as Bayesian techniques, k-nearest neighbors (kNN), the so-called Rocchio algorithm from information retrieval, artificial neural networks (ANN), support vector machines (SVM), hidden Markov models (HMMs), and decision tree (DT).

2.2. Content/feature extension for short texts

Most existing short text classification methods focus on the content/feature extension to overcome the data sparseness. One way is to expand short texts by fetching external text/knowledge. Authors in [28–30] use the short text as the query and expand the content by results returned form the search engine. Some works [31–33] exploit Wikipedia as external knowledge to enrich the short text representation with additional features. Others in [34–36] add external concepts from the knowledge base, such as Wordnet and Probase, as additional features. For feature expansion, another way is to model the latent structure of topics to connect the short text through these topics. Phan et al. [12] first build a general framework to learn classifiers with short text features combined with hidden topics. Chen et al. [38] put forward a solution by exploiting topics of multi-granularity and present a systematic way to seamlessly integrate topics and produce discriminative features for short text classification. To overcome the severe data sparsity in the short text, Cheng and Yan et al. [39] propose the biterm topic model (BTM) to learn topics by directly modeling the generation of word co-occurrence patterns (biterms) on the whole corpus.

2.3. Word vector based text representation

In the formulation of word vectors induced by language model [18–21], each word is represented by a vector which is concatenated or averaged with other word vectors in a context and the resulting vector is used to predict other words in the context. Based on word vector representations, the text level vector can be achieved in various ways. Maas et al. [14] simply used the average of all the word vectors in the document. Socher et al. [22] combined word vectors in an order given by a parse tree of a sentence, using matrix-vector operations. Le et al. [17] learned the document vector and word vectors together by concatenating the document vector with several word vectors in the document and predict the following word in the given context. All of these works learn only word vectors, while in this paper, we learn vectors of both words and topics. Compared with TWE models [42], our learning method considers more interactions between words and topics.

3. Overview of the proposed framework

In Fig. 1, we present the proposed framework consisting of three parts: topic learning, word/topic vector learning and short text classification. The topic model is estimated with LDA from the corpus, resulting in the topic model for new text inference and topic assignments of words (word-topic assignments) in each text. These word-topic assignments are used to enrich texts on the corpus for learning vectors of words and topics together. After word and topic vector learning, words and topics can be represented as vectors for text feature representation in short text classification. At the beginning of short text classification, topic inference is performed on short texts. During topic inference, each word in short texts is assigned with a topic and word-topic assignments at the end of topic inference are used to enrich each short text. Then, vectors of words and topics can be used to represent text features with more semantic information on enriched short texts. Text features are used for classifier training and new text classification. The whole framework consists of the following issues:

(1) Topic modeling on the corpus.
(2) Enriching the corpus and short texts with hidden topics.
(3) Learning vector representations of both words and topics on the enriched corpus.
(4) Feature representations of enriched short texts based on word and topic vectors.