Engineering Applications of Artificial Intelligence 64 (2017) 432–439

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Mining coherent topics in documents using word embeddings and large-scale text data

Liang Yao, Yin Zhang *, Qinfei Chen, Hongze Qian, Baogang Wei, Zhifeng Hu

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Topic model
Domain knowledge
Word embedding
Large-scale text data

ABSTRACT

Probabilistic topic models have been extensively used to extract low-dimension aspects from document collections. However, such models without any human knowledge often generate topics that are not interpretable. Recently, a number of knowledge-based topic models have been proposed, which enable users to input prior domain knowledge to produce more meaningful and coherent topics. Word embeddings, on the other hand, can automatically capture both semantic and syntactic information of words from a large amount of documents, and can be used to measure word similarities. In this paper, we incorporate word embeddings obtained from a large number of domains into topic modeling. By combining Latent Dirichlet Allocation, a widely used topic model with Skip-Gram, a well-known framework for learning word vectors, we improve the semantic coherence significantly. Our evaluation results using product review documents from 100 domains will demonstrate the effectiveness of our method.

1. Introduction

The explosive growth of online text content, such as Twitter messages, blogs, news and product reviews has brought about the challenge to understand the very dynamic sea of text. To deal with the challenge, we need to discover concepts from massive text.

A number of text mining tasks, especially aspects extraction tasks, utilize probabilistic topic models such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) (Hofmann, 1999; Blei et al., 2003). However, these unsupervised models without any human knowledge often result in topics that are difficult to interpret. In other words, they could not produce semantically coherent concepts (Chang et al., 2009; Mimno et al., 2011).

To overcome the shortcoming of interpretability in topic models, especially in LDA, some previous works incorporate prior domain knowledge into topic modeling in different ways. However, they either cannot learn knowledge automatically, or fail to utilize multiple domain data sufficiently.

Topic models such as LDA utilize the bag of word representation and document-level word co-occurrence to assign a topic to each word observation in the corpus. Similarly, word embeddings (Bengio et al., 2003; Mnih and Hinton, 2007; Collobert and Weston, 2008; Collobert et al., 2011; Huang et al., 2012; Mikolov et al., 2013) conduct dimensionality reduction based on co-occurrence information, but focus more on local context and word order in text to learn a low-dimension dense word vector for each word. Word embeddings aim at explicitly encoding many semantic relationships as well as linguistic regularities and patterns into new embedding space. For example, the result of a vector calculation vec(“Madrid”) − vec(“Spain”) + vec(“France”) is closer to vec(“Paris”) than to any other word vector, and “Spain” is close to “France” in the embedding space. Since similar words are close in embedding space, we can utilize the word correlation knowledge encoded by word embeddings.

In this paper, we improve previous knowledge-based topic models by proposing a new probabilistic method, called Word Embedding LDA (WE-LDA), which combines topic model and word embeddings, in particular LDA model and Skip-Gram (Mikolov et al., 2013). The proposed method explicitly models document-level word co-occurrence in the corpus with word correlation knowledge encoded by word vectors automatically learned from a large amount of relevant data, which could extract more coherent topics in documents.

The contributions of the paper are threefold: (1) It proposes a novel knowledge mining method for topic modeling based on word embeddings. (2) It provides a novel knowledge-based topic model which could handle the knowledge encoded by word embeddings properly.

* Corresponding author.
E-mail addresses: yaoliang@zju.edu.cn (L. Yao), yinzh@zju.edu.cn (Y. Zhang), chenqinfei@zju.edu.cn (Q. Chen), azureqianhz@zju.edu.cn (H. Qian), wbg@zju.edu.cn (B. Wei), huyangc@zju.edu.cn (Z. Hu).

http://dx.doi.org/10.1016/j.engappai.2017.06.024
Received 14 June 2016; Received in revised form 27 January 2017; Accepted 26 June 2017
0952-1976/© 2017 Elsevier Ltd. All rights reserved.

(3) Comprehensive experimental results on two large e-commerce domain datasets demonstrate our method outperforms six state-of-the-art knowledge-based topic models.

We begin this paper by introducing some related works, including studies which devote to improving the semantic coherence of topic models mainly by incorporating domain knowledge into topic models, studies which measure the coherence of topic models, and studies which focus on learning word representation. In the remainder of this paper, we first describe our model, then empirically evaluate our method on real world datasets and analyze experimental results. Experiments on two large product review datasets show the effectiveness of our method.

2. Related work

2.1. Knowledge-based topic models

To overcome the drawback of interpretability in topic models, especially in LDA, some previous works incorporate prior domain knowledge into topic modeling. For instance, Andrzejewski and Zhu (2009) proposed topic-in-set knowledge which restricts topic assignment of words to a subset of topics. Andrzejewski et al. (2011) extended topic-in-set knowledge (Andrzejewski and Zhu, 2009) by incorporating general knowledge specified by first-order logic. Similarly, Chemudugunta et al. (2008) proposed Concept model by utilizing ontologies like Open Directory Project (ODP) or The Cambridge International Dictionary of English (CIDE). The DF-LDA (Dirichlet Forest LDA) model in Andrzejewski et al. (2009) could incorporate knowledge in the form of must-links and cannot-links input by users. A must-link states that two words should share the same topic, while a cannot-link means two words should not be in the same topic. Newman et al. (2011) proposed two Bayesian regularization formulations to improve topic coherence. Both methods use additional word co-occurrence data to improve the coherence and interpretability of learned topics. Hu et al. (2011) developed a framework for allowing users to iteratively refine the topics by adding constraints that enforce that sets of words must appear together in the same topic.

Recently, LDA with Multi-Domain Knowledge (MDK-LDA) (Chen et al., 2013c) was presented, MDK-LDA is capable of using prior knowledge from multiple domains. In Chen et al. (2013b), a knowledge-based topic model, called MC-LDA (LDA with must-link set and cannot-link set), was proposed as an extension of MDK-LDA. MC-LDA assumes that all knowledge is correct and uses must-link and cannot-link knowledge as DF-LDA. GK-LDA (General Knowledge based LDA) is another knowledge-based topic model that uses the ratio of word probabilities under each topic to reduce the effect of wrong knowledge (Chen et al., 2013a). More recently, Probase-LDA (Yao et al., 2015) and Concept over Time (Yao et al., 2016), a method that combines topic model and Probase (a probabilistic knowledge base) and a method that combines topic model and Wikipedia knowledge, were put forward. The methods can model text content with the consideration of probabilistic knowledge or encyclopedia knowledge for detecting better topics. Yang et al. (2015) explored leveraging existing prior knowledge into topic modeling when dealing with large datasets.

Although above mentioned knowledge-based topic models utilize knowledge in many ways, they only process human input knowledge, but could not learn knowledge automatically.

To address the issue, AKL (Automated Knowledge LDA), LTM (Lifelong Topic model) and AMC (topic modeling with Automatically generated Must-links and Cannot-links) were proposed (Chen et al., 2014; Chen and Liu, 2014b, a), which learn knowledge automatically from multiple domains to improve topics in each domain. Our work is closely related to these methods. Despite the fact that these methods are effective, they only use frequent itemset mining to mine knowledge from top topical words in multiple domains, and do not consider the order of words in text, our method, on the other hand, utilizes multiple domain data more sufficiently by exploiting word vectors.

In order to measure the interpretability of topic models, Mimno et al. (2011) introduced an automatic coherence measure using word co-occurrence in the training corpus, which automates the human judgment approach in Chang et al. (2009). At the same time, they proposed an unsupervised generalized Pólya urn (GPU) method which improves coherence score by considering the word co-document frequency in the corpus. Newman et al. (2010) showed that an automated evaluation metric based on word co-occurrence statistics gathered from Wikipedia can reflect human evaluations of topic quality. Chuang et al. (2013) measured the correspondence between a set of latent topics and a set of reference concepts when applying topic models to domain-specific tasks, which gives another way to appraise the coherence. Word embedding has also been used to evaluate the coherence of topics from Twitter data recently Fang et al. (2016).

2.2. Word embeddings

Since the innovative work of the neural network language model (Bengio et al., 2003), a number of studies (Mnih and Hinton, 2007; Collobert and Weston, 2008; Collobert et al., 2011; Huang et al., 2012) have devoted to building distributed word representations. The dense real-valued vector representations capture local context information of words and represent their “meaning”, where the meaning of a word is determined by its surrounding words. Recently, the efficient continuous bag of words (CBOW) model and the continuous Skip-gram model (Mikolov et al., 2013) have been proposed. The training objective of CBOW is to combine the embeddings of surrounding words to predict the central word in a sliding window; while Skip-Gram tries to use the central word as input to predict the surrounding words in a sliding window.

Serving as invaluable features, word vectors have been used in NLP applications such as Part-Of-Speech tagging, chunking, named entity recognition and synonym detection (Collobert et al., 2011; Baroni et al., 2014). Word embeddings are useful because they can encode both syntactic and semantic information of words into continuous vectors and similar words are close in vector space.

Some previous studies have used word embeddings to encode semantic regularities. Nguyen et al. (2015) extended topic models by incorporating latent feature vectors of words learned from very large corpora. Das et al. (2015) replaced LDA’s parameterization of “topics” as multinomial distributions over words with multivariate Gaussian distributions on the word embedding space. However, they consider all word vectors under a topic and some less related words of a target word may result in some noise in the learning process. Moreover, learning latent feature vector and estimating parameters of multivariate Gaussian distributions are complex. Our method, on the other hand, simply uses the most related words of a target word to generate word correlation knowledge and affect the sufficient statistics directly, which is easier for model inference.

3. The WE-LDA model

The proposed WE-LDA model consists of three steps. First, we run LDA and select topical words as seed words of a corpus. Then we use word vectors to generate the must-link knowledge base. Finally, we take the generalized Pólya urn (GPU) method (Mahmoud, 2008; Mimno et al., 2011) which is the key technique for incorporating must-links into Gibbs sampling, and find more semantically coherent topics. The first and the second step aim to generate high quality prior knowledge for topic modeling. The third step is to use the learned knowledge.