Information Processing and Management 54 (2018) 593–608

Contents lists available at ScienceDirect

Information Processing and Management

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

Using semantic similarity to reduce wrong labels in distant supervision for relation extraction

Chengsen Ru, Jintao Tang, Shasha Li, Songxian Xie, Ting Wang*

College of Computer, National University of Defense Technology, No.137 Yanwachi Street, Changsha, Hunan 410073, China

ARTICLE INFO

Keywords:
Distant supervision
Relation extraction
Wrong label
Semantic Jaccard
CNN

ABSTRACT

Distant supervision (DS) has the advantage of automatically generating large amounts of labelled training data and has been widely used for relation extraction. However, there are usually many wrong labels in the automatically labelled data in distant supervision (Riedel, Yao, & McCallum, 2010). This paper presents a novel method to reduce the wrong labels. The proposed method uses the semantic Jaccard with word embedding to measure the semantic similarity between the relation phrase in the knowledge base and the dependency phrases between two entities in a sentence to filter the wrong labels. In the process of reducing wrong labels, the semantic Jaccard algorithm selects a core dependency phrase to represent the candidate relation in a sentence, which can capture features for relation classification and avoid the negative impact from irrelevant term sequences that previous neural network models of relation extraction often suffer. In the process of relation classification, the core dependency phrases are also used as the input of a convolutional neural network (CNN) for relation classification. The experimental results show that compared with the methods using original DS data, the methods using filtered DS data performed much better in relation extraction. It indicates that the semantic similarity based method is effective in reducing wrong labels. The relation extraction performance of the CNN model using the core dependency phrases as input is the best of all, which indicates that using the core dependency phrases as input of CNN is enough to capture the features for relation classification and could avoid negative impact from irrelevant terms.

1. Introduction

Relation extraction task can be defined as follows: given a sentence S with a pair of entities $e1$ and $e2$, we aim to identify the relationship between $e1$ and $e2$ (Hendrickx et al., 2010). Relation extraction is a crucial step towards natural language understanding applications, i.e. question answering (Hazrina, Sharef, Ibrahim, Murad, & Noah, 2017) and knowledge graph (Franco-Salvador, Rosso, & y Gmez, 2016; Voskarides, Meij, Tsagkias, De Rijke, & Weerkamp, 2015). However, the methods of relation extraction often encounter problems with a lack of labelled data. It is time consuming to label training data manually. To save human effort, distant supervision (DS) is firstly used by Mintz, Bills, Snow, and Jurafsky (2009) for relation extraction, which can generate labelled training data automatically. It assumes that a sentence containing an entity pair in a knowledge base expresses the corresponding relation in the knowledge base. Under this assumption, a labelled training set can be automatically generated by checking the corpus to find all the sentences containing entity pairs of the known relations. This method has been a popular choice in relation extraction for saving human effort in labelling training data (Han & Sun, 2014; Hoffmann, Zhang, Ling, Zettlemoyer, & Weld, 2011; Min, Grishman, Wan,

* Corresponding author.
E-mail address: tingwang@nudt.edu.cn (T. Wang).

https://doi.org/10.1016/j.ipm.2018.04.002
Received 12 April 2017; Received in revised form 13 March 2018; Accepted 1 April 2018
0306-4573/ © 2018 Elsevier Ltd. All rights reserved.

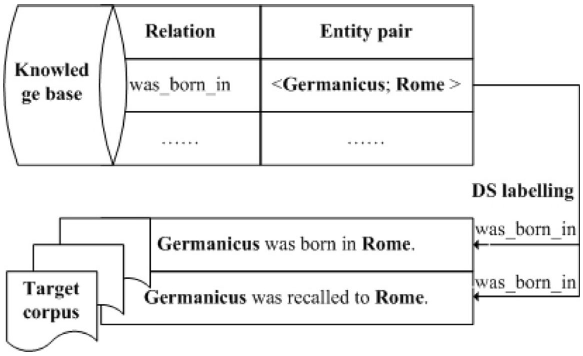

Fig. 1. The process of distant supervision.

Wang, & Gondek, 2013; Mintz et al., 2009; Riedel, Yao, & McCallum, 2010; Ritter, Zettlemoyer, Etzioni et al., 2013; Takamatsu, Sato, & Nakagawa, 2012).

However, the DS assumption may fail when there is more than one relation between an entity pair, which results in wrong labels. As shown in Fig. 1, both sentences contain the entity pair < *Germanicus, Rome* > , but express different relations, one expresses the relation *wasbornin*, the other expresses the relation *wasrecalledto*. In DS, both of them will be labelled with the same label of relation *wasbornin* or *wasrecalledto*. We have investigated the percentage of wrong labels introduced by DS in a real corpus which comes from a subset of Wikipedia containing 800,000 pages. The average error rate is 74.1% which may seriously affect the training of relation extraction models. If the wrong labels can be removed from the training data, it is expected to greatly improve the performance of relation extraction.

In the above example, we can see that the relation phrase of < *Germanicus, Rome* > in knowledge base is *wasbornin*, and the dependency phrases on the dependency path between *Germanicas* and *Rome* in two sentences are *wasbornin* and *wasrecalledto* respectively. Obviously, the dependency phrases express the relation between *Germanicas* and *Rome*.

In most cases, the relation between two entities is described by the dependency phrases between them (Vo & Bagheri, 2017). Knowledge bases use relation phrases to describe various relations. Thus, it is possible to judge whether a sentence is correctly labelled by measuring the semantic similarity between the relation phrases and the dependency phrases. The higher the semantic similarity is, the greater the probability of correct labelling tends to be. Under this assumption, a semantic similarity based method is proposed to reduce the wrong labels in DS. The semantic Jaccard algorithm is used to calculate the semantic similarity between text fragments and has been proved effective(Zhang, 2016). In this paper, the semantic Jaccard is proposed to calculate the semantic similarity with the assumption of treating the relation phrases as a text fragment and all dependency phrases belonging to the same sentence as another text fragment. When the entity pair has a semantic Jaccard value greater than a certain similarity threshold, the sentence containing the entity pair is taken as true positive.

Besides the quality of training data, the performance of traditional relation extraction methods also heavily depends on the quality of the designed features which relies on the human ingenuity and prior NLP knowledge. In recent years, some neural network models have been developed for relation extraction as an automatic feature learning method and have been proved effective (Xu, Feng, Huang, & Zhao, 2015; Zeng et al., 2014). However, these methods all suffer from irrelevant term sequences as they pay little attention to context selection and need to set the window size of context terms, which is very difficult. In the process of reducing wrong labels, the semantic Jaccard would select a core dependency phrase to represent the given relation between the entities in a sentence, which can capture the feature for relation classification without setting the window size of context words and filter irrelevant terms. Therefore, we introduce a method which uses the core dependency phrases as input to a convolution neural network (CNN) model for relation classification.

The major contributions of the work presented in this paper are as follows:

1. We propose a semantic similarity based method to reduce wrong labels appearing in DS. The method uses the semantic Jaccard to measure the semantic similarity between the relation phrase in the knowledge base and the dependency phrases between the two entities. In the process of measuring similarity, the semantic Jaccard selects the core dependency phrase to represent the given relation between the entities in a sentence.
2. We propose to take the core dependency phrases as the input of a convolutional neural network (CNN) for relation classification. The core dependency phrases can exactly capture syntactic dependency context without setting the window size of context terms and avoid negative impact from irrelevant terms.

The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 introduces the target corpus and the knowledge base. Section 4 defines the Semantic Jaccard. Section 5 gives the objective and framework of this paper. Section 6 introduces the method to use the semantic Jaccard to reduce wrong labels and describes the CNN model for relation classification. Section 7 evaluates the performance of the proposed method. Section 8 gives the conclusion.

Download English Version:

https://daneshyari.com/en/article/6926002

Download Persian Version:

https://daneshyari.com/article/6926002

Daneshyari.com