Engineering Applications of Artificial Intelligence 36 (2014) 195–203

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim*, Ali Selamat

Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 5 April 2014
Received in revised form
15 July 2014
Accepted 25 July 2014
Available online 24 August 2014

Keywords:
Cross-lingual
Sentiment classification
Multiple source languages
Multi-view learning
Semi-supervised learning

ABSTRACT

Cross-lingual sentiment classification aims to utilize annotated sentiment resources in one language (typically English) for sentiment classification of text documents in another language. Most existing research works rely on automatic machine translation services to directly project information from one language to another. However, due to the existence of differing linguistic terms and writing styles between different languages, translated data cannot cover all vocabularies which exist in the original data. Further, different term distribution between translated data and original data can lead to low performance in cross-lingual sentiment classification. To overcome these problems, we propose a new model which uses labelled data from multiple source languages in a multi-view semi-supervised learning approach so as to incorporate unlabelled data from the target language into the learning process. The proposed model was applied to book review datasets in four different languages. Experiments have shown that our model can effectively improve the cross-lingual sentiment classification performance in comparison with some baseline methods.

1. Introduction

Along with the rapid increase of internet access in the world today, the volume of user-generated content on the web has also intensified. Due to the enormous amount of this content, the task of summarizing their information into a useful format is a very difficult and challenging problem. This challenge has motivated the natural language processing (NLP) communities to design and develop computational methods by which to mine and analyze the information of these text documents. Opinion mining or sentiment analysis is one of the most interesting fields in this area; it analyzes people's opinions, attitudes and sentiments towards entities such as products, individuals, events, etc. (Liu, 2012). Text sentiment classification refers to the task of determining the sentiment polarity (e.g. positive or negative) of a given text document and has received considerable attention due to its many useful applications in product review classification (Zhou et al., 2013) and opinion summarization (Ku et al., 2006).

Up until now, different methods have been used in sentiment classification. These methods can be categorized into two main groups, namely: lexicon-based and corpus-based. The lexicon-based methods classify text documents based on the polarity of words and phrases contained in the text (Taboada et al., 2011; Turney, 2002). This group of methods requires a sentiment lexicon to distinguish between positive and negative terms. In contrast, corpus-based methods train a sentiment classifier based on a labelled corpus using machine learning classification algorithms (Moraes et al., 2013; Pang et al., 2002). The performance of these methods depends intensively on both the quantity and quality of labelled corpus items as the training set.

Sentiment lexicons and annotated sentiment corpora are the most important resources for sentiment classification. However, since most recent research studies in sentiment classification are written in a limited number of languages (always English), this has led to a scarcity of labelled corpus and sentiment lexicons in other languages (Martín-Valdivia et al., 2013; Wan, 2011). Further, manual construction of reliable sentiment resources is a very difficult and time-consuming task. Therefore, the challenge is how to utilize labelled sentiment resources in one language (i.e. English) for sentiment classification into another language. This leads to an interesting research area called cross-lingual sentiment classification (CLSC). The most direct solution to this problem is the use of machine translation systems to directly project the information of data from one language into another (Balahur et al., 2014; Banea et al., 2008; Martín-Valdivia et al., 2013; Prettenhofer and Stein, 2010; Wan, 2009, 2011). Most existing works in this area have used machine translation systems to translate labelled training data from the source language into the target language and perform sentiment classification into the target language (Balahur and Turchi, 2014; Banea et al., 2010). Some other

* Corresponding author. Tel.: +60 7 5538727.
E-mail address: roliana@utm.my (R. Ibrahim).

http://dx.doi.org/10.1016/j.engappai.2014.07.020
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

researchers have employed machine translation in the opposite direction to translate unlabelled test data from the target language into the source language and to perform the classification in the source language (Hajmohammadi et al., 2014b; Martín-Valdivia et al., 2013; Prettenhofer and Stein, 2010). A limited number of research works have used both directions of translation to create two different views of the training and test data to compensate for some of the translation limitations (Hajmohammadi et al., 2014a; Pan et al., 2011; Wan, 2009, 2011).

However, because the training set and the test set come from two different languages having differing linguistic terms and writing styles, as well as originating from different cultures, translated text documents cannot cover all the vocabularies contained in the original text documents. Therefore, these methods cannot attain the performance results of monolingual sentiment classification methods in which the training and test samples are from the same language. Using multiple resources from multiple languages can alleviate the problem of vocabulary coverage in CLSC. This occurs because some vocabularies which are not covered by the feature set extracted from translated documents of the one source language may be covered by the feature set of another source language. This means that feature sets extracted from the training data of multiple source languages can cover more vocabularies of test documents in the target language. For example, the translation of “awesome” into French is “génial” but a word in German with the same meaning “fantastisch” is translated to “fantastique” in French. Both words “génial” and “fantastique” are used in the French reviews and each word is covered by a different source language. Therefore, using a multiple source language technique is expected to show better performance in CLSC when compared with models which use only one source language.

Different term distribution between the original and the translated text documents is another important factor that can reduce the performance of CLSC. It means that a term may be frequently used in one language to express an opinion while the translation of that term is rarely used in the other language. To overcome this problem, making use of unlabelled data from the target language can be helpful, since this type of data is always easy to obtain and has the same term distribution as the test data. Therefore, employing unlabelled data from the target language in the learning process is expected to result in better classification in CLSC.

In the light of difficulties for CLSC, we address the task of CLSC via a multi-view semi-supervised learning framework. Specifically, we propose a new learning model that uses labelled data from multiple languages (in this paper, two languages) as multiple training datasets. Both directions of translation are then used to create different views of data. These individual views are then employed in a multi-view semi-supervised learning process to incorporate unlabelled data from the target language in the learning process.

The contributions of our work are as follows: (1) utilizing training data and their translations from multiple source languages in CLSC to cover more vocabularies of test documents in the target language; (2) employing a multi-view semi-supervised learning strategy in order to incorporate unlabelled examples from the target language in the learning process of CLSC. This is achieved by creating multiple views from the documents in both the source and the target languages through automatic machine translation and using the “majority teaching minority” strategy to select the most confident pseudo-labelled examples from unlabelled documents and adding them to the training sets in each of the individual views.

The proposed model was applied to book review datasets in four different languages. Experiments showed that the use of this model obtained better performance in comparison with some baseline methods.

The reminder of this paper is organized as follows: the next section presents related works on CLSC. Section 3 describes multiple views data creation. The proposed model is described in Section 4, while an evaluation is given in Section 5. Finally, Section 6 concludes this paper and outlines ideas for future research.

2. Related works

Cross-lingual sentiment classification has been extensively studied in recent years. These research studies are based on the use of annotated data in the source language (always English) to compensate for the lack of labelled data in the target language. Most approaches focus on resource adaptation from one language to another with few sentiment resources. For example, Mihalcea et al. (2007) generate subjectivity analysis resources into a new language from English sentiment resources by using a bilingual dictionary. In other works (Banea et al., 2010; Banea et al., 2008), automatic machine translation engines were used to translate the English resources for subjectivity analysis. In a further study (Banea et al., 2008), the authors showed that automatic machine translation is a viable alternative to the construction of resources for subjectivity analysis in a new language. In two different experiments, they first translated training data of subjectivity classification from the source language into the target language. They then utilized this translated data to train a classifier in the target language and applied this trained classifier to classify test data. Additionally, in another experiment, machine translation was used to translate test data from the target language into the source language and a classifier was then trained based on training data in the source language. After the training phase, the translated test data was presented to the classifier for sentiment polarity prediction. Wan (2008) used unsupervised sentiment polarity classification in Chinese product reviews. He translated Chinese reviews into different English reviews using a variety of machine translation engines and then performed sentiment analysis for both the Chinese and English reviews using the lexicon-based technique. Finally, he used ensemble methods by which to combine the analysis results. This method requires sentiment lexicon in the target language and cannot be applied to other languages with no lexicon resource. Pan et al. (2011) designed a bi-view non-negative matrix tri-factorization (BNMTF) model in an attempt to solve the problem of cross-lingual sentiment classification. They used the machine translation to achieve two representations of training and test data; one in the source language and another in the target language. This model was then used to combine the information from two views.

Another approach is that of feature translation, which involves translating the features extracted from labelled documents (Moh and Zhang, 2012; Shi et al., 2010). The features, selected by a feature selection algorithm, are translated into different languages. Subsequently, based on those translated features, a new model is trained for each language. This approach only needs a bilingual dictionary to translate the selected features. It can, however, suffer from the inaccuracies of dictionary translation, in that words may have different meanings in different contexts. Therefore, selecting the features to be translated can be an intricate process. Prettenhofer and Stein (2010) investigated CLSC from the domain adaptation view by employing structural correspondence learning (SCL) (Blitzer et al., 2006). They adapted SCL to use unlabelled data and a word translation oracle to induce correspondence among the words from both the source and target languages. They first selected some word pairs called pivots and then identified correlations between pivots and other words in unlabelled documents. After that, a map was extracted to associate the original representation of a document in the source and target languages