Information Processing and Management 50 (2014) 718–732

Contents lists available at ScienceDirect

Information Processing and Management

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

Bi-view semi-supervised active learning for cross-lingual sentiment classification

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim *, Ali Selamat

Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81300 UTM Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 7 June 2013
Received in revised form 14 March 2014
Accepted 17 March 2014
Available online 24 April 2014

Keywords:
Cross-lingual
Sentiment classification
Co-training
Active learning
Density measure

ABSTRACT

Recently, sentiment classification has received considerable attention within the natural language processing research community. However, since most recent works regarding sentiment classification have been done in the English language, there are accordingly not enough sentiment resources in other languages. Manual construction of reliable sentiment resources is a very difficult and time-consuming task. 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, different term distribution between original and translated text documents and translation errors are two main problems faced in the case of using only machine translation. To overcome these problems, we propose a novel learning model based on active learning and semi-supervised co-training to incorporate unlabelled data from the target language into the learning process in a bi-view framework. This model attempts to enrich training data by adding the most confident automatically-labelled examples, as well as a few of the most informative manually-labelled examples from unlabelled data in an iterative process. Further, in this model, we consider the density of unlabelled data so as to select more representative unlabelled examples in order to avoid outlier selection in active learning. The proposed model was applied to book review datasets in three different languages. Experiments showed that our model can effectively improve the cross-lingual sentiment classification performance and reduce labelling efforts in comparison with some baseline methods.

1. Introduction

Text sentiment classification refers to the task of determining the sentiment polarity (e.g. positive or negative) of a given text document (Liu, 2012). Recently, sentiment classification has received considerable attention in the natural language processing research community due to its many useful applications such as online product review classification (Kang, Yoo, & Han, 2012) and opinion summarization (Ku, Liang, & Chen, 2006).

Up until now, different methods have been used for sentiment classification. These methods can be categorised into two groups, namely; unsupervised and supervised. The unsupervised methods classify text documents based on the polarity of words and phrases contained in the text. If a text document contains more positive than negative terms, for example, it is

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

http://dx.doi.org/10.1016/j.ipm.2014.03.005
0306-4573/© 2014 Elsevier Ltd. All rights reserved.

classified as positive and vice versa (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Turney, 2002). A sentiment lexicon is always used to determine the sentiment polarity of each term. In contrast, supervised methods train a sentiment classifier based on labelled data using some machine learning classification algorithms (Pang, Lee, & Vaithyanathan, 2002; Ye, Zhang, & Law, 2009). The performance of these methods specifically depends on the quality of labelled data as a training set for the sentiment classifier.

Based on these two groups of methods, sentiment lexicons and annotated sentiment data can be seen as the most important resources for sentiment classification. However, since most recent research studies in sentiment classification have been presented in the English language, there are not enough labelled corpus and sentiment lexicons in other languages (Montoyo, Martínez-Barco, & Balahur, 2012). 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 (a resource-rich language such as English is always called the source language) for sentiment classification in another language (a resource-scarce language is called the target language). This subsequently leads to an interesting area of research called cross-lingual sentiment classification (CLSC). Most existing research works employ automatic machine translation engines to directly project information of labelled data from the source language into the target language (Balahur & Turchi, 2014; Banea, Mihalcea, & Wiebe, 2010). In this case, a sentiment classifier is trained based on the translated labelled data and then applied to the original test data for the classification task in the target language. Machine translation can be employed in the opposite direction by translating test documents from the target language into the source language (Martín-Valdivia, Martínez-Cámara, Perea-Ortega, & Ureña-López, 2013; Prettenhofer & Stein, 2011). In this situation, the sentiment classifier is trained based on the original labelled data in the source language and then applied to the translated test data. However, the use of only translated data in the sentiment classification task results in two main problems. The first problem is the difference in term distribution between the original and the translated text documents due to the dissimilarity in cultures, writing styles and also linguistic expressions in the various languages. This subsequently leads to the creation of different feature distributions in the training and test data. The second problem relates to machine translation errors in the resource translation process. However, since machine translation quality is still far from satisfactory, there are some translation errors which occur in the resource projection process. To overcome the first problem, making use of unlabelled data from the target language can be helpful, because 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 a better classification in CLSC.

Active learning (AL) (Wang, Kwong, & Chen, 2012) and semi-supervised learning (SSL) (Ortigosa-Hernández et al., 2012) are two well-known techniques that make use of unlabelled data to improve classification performance. Both techniques are iterative processes. AL aims to reduce manual labelling efforts by finding the most informative examples for human labelling, while SSL tries to automatically label examples from unlabelled data in each cycle.

To reduce the effect of machine translation errors in the classification process, both directions of machine translation can be simultaneously employed. Therefore, we have training and test documents in both languages (original version of training documents and translated version of test documents in the source language and translated version of training documents and original version of test documents in the target language). If the translated version of a document has some translation error on one side, the original version of that document is used on the other side.

Given the two possible directions for data translation, we can consider sentiment data from two different views, namely; source language view and target language view. In this paper, we consider source language features and target language features as being two sufficient feature representations of labelled and unlabelled data. Accordingly, we propose a new model based on a combination of bi-view Active learning, Co-testing (Muslea, Minton, & Knoblock, 2006), and semi-supervised co-training (Park & Zhang, 2004) in order to incorporate unlabelled data from the target language into the learning process. Co-testing is a bi-view active learning process that aims to find the most informative unlabelled examples by considering the disagreement between two classifiers trained in each view. The intuitive theory behind co-testing is that if the classifiers trained in each view classify an unlabelled example differently, at least one classifier makes a mistake on its prediction. Therefore, this unlabelled example can provide useful information for the classifier with an incorrect prediction. On the other hand, Co-training tries to select the most confidently-classified examples from unlabelled data in each view so as to add to the training data in the other view. These two techniques complement each other in order to reduce human labelling efforts.

From experimental results in co-testing, it can be seen that some of the selected contention examples cannot provide much help to the learner. The main reason for this issue is that some of these selected examples are outliers and therefore are not representatives. To avoid outlier selection in co-testing, we considered the density of selected examples in our proposed method and chose those contention examples that have maximum average density in the pool of unlabelled data.

The contributions of our work are as follows: (1) parallel combining of two bi-view approaches, co-training and co-testing, in order to incorporate unlabelled examples from the target language in the learning process of cross-lingual sentiment classification. This is achieved by selecting the most confident automatically-labelled examples, as well as a few of the most informative manually-labelled examples from unlabelled data. Specifically, the contribution degree was defined as the criteria of select informative instances. These select the contention examples which have a different predicted label between the source and target views of unlabelled documents. (2) When selecting the most informative unlabelled examples by co-testing, we propose to use density information of unlabelled examples in order to select not only the most informative examples, but also the most representative examples for manual labelling. Specifically, we select the contention