Graph-Based Semi-supervised Learning for Cross-Lingual Sentiment Classification

Mohammad Sadegh Hajmohammadi[1(✉)], Roliana Ibrahim[2], and Ali Selamat[2]

[1] Department of Computer Engineering, Sirjan Branch, Islamic Azad University, Sirjan, Iran
hajmohammadi@iausirjan.ac.ir
[2] Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
{roliana,aselamat}@utm.my

Abstract. Cross-lingual sentiment classification aims to use labelled sentiment data in one language for sentiment classification of text documents in another language. Most existing research works rely on automatic machine translation services to directly transfer information from one language to another. However, different term distribution between translated data and original data can lead to low performance in cross-lingual sentiment classification. Further, due to the existence of differing structures and writing styles between different languages, using only information of labelled data from a different language cannot show a good performance in this classification task. To overcome these problems, we propose a new model which uses sentiment information of unlabelled data as well as labelled data in a graph-based semi-supervised learning approach so as to incorporate intrinsic structure of unlabelled data from the target language into the learning process. The proposed model was applied to book review datasets in two different languages. Experiments have shown that our model can effectively improve the cross-lingual sentiment classification performance in comparison with some baseline methods.

Keywords: Cross-lingual · Sentiment classification · Graph-based · Semi-supervised learning

1 Introduction

Text sentiment classification refers to the task of determining the sentiment polarity (e.g. positive or negative) of a given text document [1]. Recently, sentiment classification has received considerable attention in the natural language processing research community due to its many useful applications such as opinion summarization [2] and online product review classification [3].

Up until now, different approaches have been employed in sentiment classification. These approaches can be divided into two main groups, namely; unsupervised and supervised methods. The unsupervised methods classify text documents based on the polarity of words and phrases contained in the text [4, 5]. This group of methods needs a sentiment lexicon to distinguish between the positive and negative terms. In contrast,

N.T. Nguyen et al. (Eds.): ACIIDS 2015, Part I, LNAI 9011, pp. 97–106, 2015.
DOI: 10.1007/978-3-319-15702-3_10

supervised methods train a sentiment classifier based on labelled corpus using machine learning classification algorithms [6, 7]. The performance of these methods intensively depends on the quantity and the quality of labelled corpus as the training set.

Based on these two groups of methods, sentiment lexicons and annotated sentiment corpora 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 [8]. 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 for sentiment classification in another language. This subsequently leads to an interesting research area called cross-lingual sentiment classification (CLSC).

The most direct solution of this problem is the use of machine translation systems to directly project the information of data from one language into the other language [9-12]. The most existing research works develop a sentiment classifier based on the translated labelled data from the source language and use this classifier to determine the sentiment polarity of test data in the target language [13, 14]. Machine translation can be employed in the opposite direction by translating the test documents from the target language into the source language [15, 16]. 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. A few number of research works used both direction of translation to create two different views of the training and the test data to compensate some of the translation limitations [9, 10]. But because the training set and the test set are from two different languages with different intrinsic structures and writing styles and also originate from different cultures, these methods cannot reach the performance of monolingual sentiment classification methods in which the training and test samples are from the same language. Recently, some research works try to incorporate unlabelled document from the target language into the learning process of sentiment classification to fill the gaps between original and translated documents [9-11, 17, 18]. Although using unlabelled data from the target language can help to improve the classification performance, CLSC cannot reach the performance of mono-lingual sentiment classification because intrinsic structure of documents in the target language is fixed and different from the documents in the source language. Therefore, incorporating the intrinsic structure of documents in the target language is expected to result in better performance in CLSC. In fact, a good CLSC model should uses the information of the source language data while following the structure of the target language documents.

In this paper, a new model of CLSC is designed by taking into account the labelled documents in the source languages as well as the intrinsic structure of unlabelled documents in the target language. This model is based on the graph-based semi-supervised learning approach.

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 language with few sentiment resources. For example, Mihalcea et al.[19] generated subjectivity analysis resources into a new language from English sentiment resources by using a bilingual dictionary. Wan [20] 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 Chinese and English reviews using a lexicon-based technique. Finally, he used ensemble methods to combine the results of analysis.

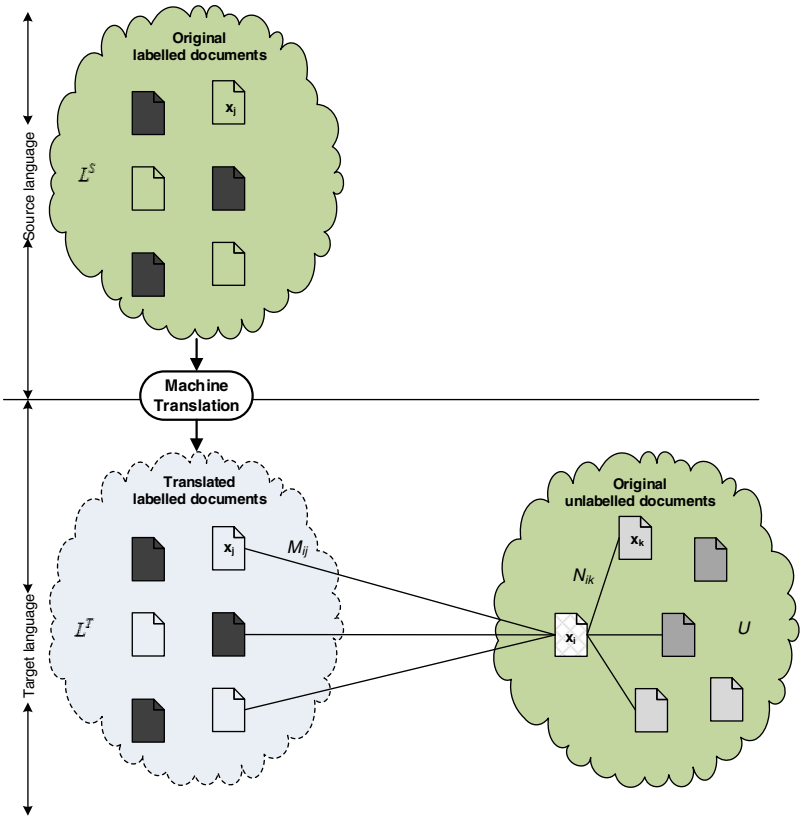

Fig. 1. Graph construction process in graph-based model

In another work, Wan [9] used the co-training method to overcome the problem of cross-lingual sentiment classification. In this paper, he exploited a bilingual co-training approach to leverage annotated English resources to sentiment classification in Chinese reviews. In this work, firstly, machine translation services were used to translate English labelled documents (training documents) into Chinese and similarly, Chinese unlabeled documents into English. The author used two different views (English and Chinese) in order to exploit the co-training approach into the classification problem. In an early work, Hajmohammadi et al. [11] tried to utilize multiple source

languages in the process of CLSC. They showed that using more source languages can help to cover more information of sentiment terms in the classification process. To the best of our knowledge, graph-based method has not yet been investigated in the field of cross-lingual sentiment classification.

3 Proposed Model

This model is designed so as to incorporate the intrinsic structure of review documents in the target language into the learning process of CLSC. For this task, two different weighted graph are constructed based the translated labelled documents and original unlabelled documents in the target language. The process of graphs construction is illustrated in Fig. 1.

At the beginning of the learning process, a sentiment score is assigned to every document in both labelled and unlabelled sets. After that, the sentiment scores of review documents in the unlabelled set are iteratively computed by using the predefined labels of translated labelled documents as well as the pseudo-labels of original unlabelled documents in the target language. This learning process can be described as follows:

1. Suppose, U denotes the unlabelled document set represented in the target language. Also suppose L^T , denotes the translated version of labelled document set represented in the target languages. Y_U denotes the sets of sentiment scores for documents in U . The sentiment score set of L^T is also represented by Y_L .
2. Traditional supervised classification is used to determine the pseudo-labels of documents in U using corresponding labelled sets, L^T . Y_U is initialized using these determined labels. The initial label of a document is set to 1, if the document is labelled “positive”, and to -1, if the document is labelled “negative”.
3. The sentiment scores in each score set are normalized such that the sum of positive scores becomes 1 and the sum of negative scores becomes -1.
4. Cosine similarity measure is used to compute the pairwise similarity values between two documents (both labelled and unlabelled documents). Each document is represented by a feature vector, each entry of which contains a feature weight. TF-IDF is used as feature weights.
5. A graph is constructed based on the labelled and unlabelled documents represented in the target language. The nodes of this graph represent documents in L^T and U . The edges of this graph represent the content similarities between documents in U and documents in L^T . A similarity matrix, M , is created from the documents in L^T and U and normalized such that the sum of each row becomes 1. The normalized matrix is sorted in descending order for every row in order to find the nearest neighbors of a document.

6. A matrix $\tilde{M}$ is used to denote the *k*-nearest neighbors of U in the labelled set. Therefore, Y_U , the sentiment scores of unlabelled documents , can be computed as follows:

$$Y_U^{(k)}(i) = \sum_{j \in \tilde{M}_i} (M_{ij} \times Y_L(\mathrm{j})) \tag{1}$$

Where $Y_U^{(k)}(i)$ represents the sentiment score of *i*th document in U at the *k*th iteration, and $Y_L(j)$ represent the sentiment score of *j*th document in L^T .

7. In the same way, a graph is constructed using only unlabelled documents. The nodes of this graph represent documents in U and the edges denote the similarities between unlabelled documents. A similarity matrix, N , is created from these similarity scores and also normalized such that the sum of each row becomes 1. The normalized matrix is sorted in descending order for every row in order to find the nearest neighbors of a document.
8. A matrix $\tilde{N}$ is used to denote the k-nearest neighbors of documents in U . Therefore, Y_U , the sentiment scores of unlabelled documents, can be computed as follows:

$$Y_U^{(k)}(i) = \sum_{j \in \tilde{N}_i} (N_{ij} \times Y_U^{(k-1)}(\mathrm{j})) \tag{2}$$

9. In order to incorporate the sentiment scores of neighbors document in both labelled and unlabelled sets, the above iterative formulas are combined and two new iterative formulas are obtained to compute Y_U as follows:

$$Y_U^{(k)} = \alpha M Y_L + \beta N Y_U^{(k-1)} \tag{3}$$

Where α and β demonstrate the relative effect of labelled and pseudo-labelled data in final sentiment score computation and α+β=1.

10. Y_U is normalized at every iteration such that the sum of positive scores becomes 1 and the sum of negative scores becomes -1. This normalization process is needed for algorithm convergence. The iterative process is continued until convergence.
11. A sentiment label is assigned to each document in unlabelled pool according to calculated sentiment scores in Y_U . If the sentiment score is in the range of 0 to +1, then the document is labelled as “positive”. If this score is between -1 and 0, then the document is labelled as “negative”.

The convergence of the algorithm occurs when the difference between the sentiment scores calculated at two consecutive steps of algorithm for all unlabelled examples falls below the certain threshold.

As described in this process, the sentiment score for each unlabelled document is calculated based on two different graphs. One graph is constructed to connect the unlabelled documents to the labelled documents and another graph is constructed to represent the inter-connection of unlabelled documents. Consequently, the sentiment score of an unlabelled document is computed by incorporating the similarities of that document to the labelled documents as well as its similarities to other pseudo-labelled (unlabelled) documents. This means that each unlabelled document receives a sentiment score from both labelled and unlabelled examples. Due to the existence of similar intrinsic structures among unlabelled documents, incorporating their sentiment scores is expected to improve the performance of CLSC in compare to other methods.

4 Evaluation

In this section, we evaluate our proposed approach in CLSC on two different languages in the book review domains and compare it with some baseline methods.

4.1 Datasets

Two different evaluation datasets have been used in this paper.

- English-Japanese dataset (En-Jp): This dataset contains Amazon book review documents in English and Japanese languages. This dataset was used by Prettenhofer and Stein [21] .
- English-Chinese dataset (En-Ch): This dataset was selected from Pan reviews dataset [18]. It contains book review documents in English and Chinese languages.

Table 1 shows the characteristics of these two datasets. All review documents in the source language (English) are translated into the target languages using the Google translate engine[1]. In the Japanese text document, we applied MeCab[2] segmenter software to segment the reviews; while Chinese documents were segmented by the Stanford Chinese word segmenter[3]. In the feature extraction step, unigram and bi-gram patterns were extracted as sentimental patterns. To reduce computational complexity, especially in density estimation, we performed feature selection using the information gain (IG) technique. We selected 5000 high score unigrams and bi-grams as final features. Each document is represented by a feature vector, each entry of which contains a feature weight. We used TF-IDF as feature weights.

[1] http://translate.google.com/
[2] http://mecab.googlecode.com/svn/trunk/mecab/
[3] http://nlp.standfor.edu/software/segmenter

Table 1. Charactristices of datasets used in the evaluation

Dataset	Domain	Languages		Total documents	Positive documents	Negative documents
En-Ch [18]	Book review	Source Language	English	2000	1000	1000
		Target Language	Chinese	4000	2000	2000
En-Jp [21]	Book review	Source Language	English	2000	1000	1000
		Target Language	Japanese	4000	2000	2000

4.2 Baseline Methods

The following baseline methods are implemented in order to evaluate the effectiveness of proposed models.

- Co-training: This is the traditional co-training algorithm which was used in the study by [9, 22]. It uses labelled data from the source language and unlabelled data from the target language in two views.
- Structural Correspondence Learning model (SCL): This model was implemented as introduced in [13]. The Google Translate service was used to map words in the source vocabulary to the corresponding translation in the target vocabulary. Other parameters were set as used in [13].
- Transductive SVM in the source language (TSVM): This method uses the well-known transductive learning model based on support vector machine (SVM) for sentiment classification. In this model a transductive SVM is trained based on the translated labelled document and original unlabelled documents.

4.3 Results and Discussion

In this section, the proposed model is compared with three baseline methods. In the proposed algorithm, α and β were set to 0.4 and 0.6 respectively, which indicates the contribution from unlabelled data is a little more important than that from labelled data. The threshold also was set to 0.1e-08 for convergence condition. The parameter k was set to 30 in k-nearest neighbor matrix. Cosine measure was used to determine the content similarity between documents.

Table 2 and Table 3 show the numerical results for comparing the proposed model and the baseline methods. As we can see in these tables, the proposed model can show a good performance in compare to all of the baseline methods and obtained the best accuracy in all datasets.

Table 2. Performance comparison in English-Japanese (En-Ch) dataset (best results are reported in bold-face type)

Methods	Accuracy	Positive			Negative		
		Pre	Rec	F1	Pre	Rec	F1
Proposed model	**73.81**	79.27	64.30	71.00	70.10	**83.27**	**76.12**
Co-Training	73.32	77.17	66.75	71.59	70.38	79.90	74.84
SCL	70.58	70.89	69.24	70.06	70.28	71.90	71.08
TSVM	71.75	71.60	**71.85**	**71.73**	**71.90**	71.64	71.77

Table 3. Performance comparison in English-Japanese (En-Jp) dataset (best results are reported in bold-face type)

Methods	Accuracy	Positive			Negative		
		Pre	Rec	F1	Pre	Rec	F1
Proposed model	**72.72**	**75.18**	67.83	**71.32**	**70.70**	**77.61**	**73.99**
Co-Training	72.27	74.54	67.73	70.96	70.40	76.80	73.45
SCL	69.50	72.89	62.39	67.23	67.26	76.60	71.63
TSVM	69.02	69.00	**69.07**	69.03	69.04	68.97	69.00

Compared to the co-training and SCL models, proposed model shows better overall accuracy in all datasets. This is due to the taking into account the intrinsic structure of documents in the target language during the sentiment scores prediction process.

In compare to transductive SVM (TSVM), the proposed model shows better performance in almost all datasets. This means that, incorporation of document similarities has a beneficial effect in the sentiment score prediction process.

5 Conclusion

In this paper, we have proposed a new graph-based semi-supervised learning model to improve the performance of cross-lingual sentiment classification. In the proposed model, automatic machine translation was used to project the information of source language documents into the target languages. Two different graphs were constructed based on the similarity measure between the labelled and unlabelled document and among unlabelled documents. The sentiment score of each unlabelled document was then computed through propagation of sentiment scores of labelled and unlabelled documents. This model was applied to the cross-lingual sentiment classification dataset in two different languages and the performance of the proposed model was compared with some baseline methods. The experimental results show that this model can improve the performance of CLSC in compare to the baseline methods.

References

1. Liu, B., Zhang, L.: A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer US(2012)
2. Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs (2006)
3. Kang, H., Yoo, S.J., Han, D.: Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst. Appl. **39**(5), 6000–6010 (2012)
4. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania: Association for Computational Linguistics (2002)
5. Taboada, M., et al.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics (2002)
7. Moraes, R., Valiati, J.F., Neto, W.P.G.: Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Syst. Appl. **40**(2), 621–633 (2013)
8. Montoyo, A., Martínez-Barco, P., Balahur, A.: Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decis. Support Syst. **53**(4), 675–679 (2012)
9. Wan, X.: Bilingual co-training for sentiment classification of Chinese product reviews. Comput. Linguist. **37**(3), 587–616 (2011)
10. Hajmohammadi, M.S., Ibrahim, R., Selamat, A.: Bi-view semi-supervised active learning for cross-lingual sentiment classification. Inf. Process. Manage. **50**(5), 718–732 (2014a)
11. Hajmohammadi, M.S., Ibrahim, R., Selamat, A.: Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning. Eng. Appl. Artif. Intell. **36**, 195–203 (2014b)
12. Balahur, A., Turchi, M.: Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language (2013)
13. Prettenhofer, P., Stein, B.: Cross-language text classification using structural correspondence learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1118–1127. Association for Computational Linguistics, Uppsala, Sweden (2010)
14. Perea-Ortega, J.M., et al.: Improving polarity classification of bilingual parallel corpora combining machine learning and semantic orientation approaches. J. Am. Soc. Inform. Sci. Technol. **64**(9), 1759–1962 (2013)
15. Banea, C., Mihalcea, R., Wiebe, J.: Multilingual subjectivity: are more languages better? In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 28–36. Association for Computational Linguistics: Beijing, China (2010)
16. Balahur, A., Turchi, M.: Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. **28**(1), 56–75 (2014)
17. Hajmohammadi, M.S., Ibrahim, R., Selamat, A.: Density based active self-training for cross-lingual sentiment classification. In: Jeong, H.Y., Obaidat, M.S., Yen, N.Y., Park, J.J. (eds.) Advanced in Computer Science and Its Applications. LNEE, vol. 279, pp. 1053–1059. Springer, Heidelberg (2014c)

18. Pan, J., Xue, G.-R., Yu, Y., Wang, Y.: Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part I. LNCS(LNAI), vol. 6634, pp. 289–300. Springer, Heidelberg (2011)
19. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (2007)
20. Wan, X.: Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 553–561. Association for Computational Linguistics, Honolulu (2008)
21. Prettenhofer, P., Stein, B.: Cross-Lingual Adaptation Using Structural Correspondence Learning. ACM Trans. Intell. Syst. Technol. **3**(1), 1–22 (2011)
22. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 235–243. Association for Computational Linguistics: Suntec, Singapore (2009)