A Mixed Model for Cross Lingual Opinion Analysis*

Lin Gui[1], Ruifeng Xu[1,**], Jun Xu[1], Li Yuan[1], Yuanlin Yao[1], Jiyun Zhou[1], Qiaoyun Qiu[1], Shuwei Wang[1], Kam-Fai Wong[2], and Ricky Cheung[3]

[1] Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
{Guilin.nlp,xuruifeng.hitsz,hit.xujun,yuanlisail}@gmail.com, {541523621,262061485,1028440296}@qq.com, wangshuweino1@vip.qq.com

[2] Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong
kfwong@se.cuhk.edu.hk

[3] Social Analytics (Hong Kong) Co. Ltd., Hong Kong
ricky.cheung@hkmci.com

Abstract. The performances of machine learning based opinion analysis systems are always puzzled by the insufficient training opinion corpus. Such problem becomes more serious for the resource-poor languages. Thus, the cross-lingual opinion analysis (CLOA) technique, which leverages opinion resources on one (source) language to another (target) language for improving the opinion analysis on target language, attracts more research interests. Currently, the transfer learning based CLOA approach sometimes falls to over fitting on single language resource, while the performance of the co-training based CLOA approach always achieves limited improvement during bi-lingual decision. Target to these problems, in this study, we propose a mixed CLOA model, which estimates the confidence of each monolingual opinion analysis system by using their training errors through bilingual transfer self-training and co-training, respectively. By using the weighted average distances between samples and classification hyper-planes as the confidence, the opinion polarity of testing samples are classified. The evaluations on NLP&CC 2013 CLOA bakeoff dataset show that this approach achieves the best performance, which outperforms transfer learning and co-training based approaches.

Keywords: Cross lingual Opinion Analysis, Transfer Self-Training, Co-Training, Mixed Model.

* This research is supported by Open Projects Program of National Laboratory of Pattern Recognition, National Natural Science Foundation of China No. 61203378, 61370165, MOE Specialized Research Fund for the Doctoral Program of Higher Education 20122302120070, Shenzhen Foundational Research Funding JCYJ20120613152557576 and Shenzhen International Cooperation Research Funding GJHZ2012 0613110641217.

** Corresponing author.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 93–104, 2013.

1 Introduction

With the rapid development of Internet, mass texts are published on self-media, such as micro-blogging, news reviews, product reviews, etc. These texts contain many valuable subjective opinions. Thus, the opinion analysis technique, which identifies and analyzes the opinions in these text become a focus topic in natural language processing research. Generally speaking, current opinion analysis techniques are camped into rule-based or machine learning based approaches. Rule-based opinion analysis approach achieves good performances in some specific tasks, but it requires manually compile of rules and patterns. The machine learning based, especially supervised learning based opinion analysis approach is the majority of existing techniques. However, their performances are always influenced by the insufficient training corpus. Such problem becomes more serious for some resource-poor languages. Thus, the cross-lingual opinion analysis (CLOA) techniques are investigated. The main idea of CLOA is to translate and project the opinion resources from one language (named as source language) to another language (named as target language) for enriching target language resources, and thereby improving the performance of opinion analysis on target language.

Normally, there are much noises occurred during mapping and projecting the source language resources to target language, which affects the classifier training. Therefore, one of the core problems in CLOA is to investigate effective mechanism for selecting high quality transferred samples for classifier training. Most existing CLOA techniques in general can be divided into two categories: 1. Transfer learning based approach, which transfers the source language resources to target language, including dictionary and opinion corpus, for improving the classifier on target language; and 2.Co-training based approach, which adopts co-training strategy to improve the opinion analysis performance on both languages. A typical transfer learning based CLOA works is conducted by J. Wiebe [1], which utilized English and Romanian aligned corpus to generate a Romanian subjectivity dictionary and then developed a dictionary/ruled based Romanian opinion analysis system. Yao [2] used Chinese and English cross-lingual dictionary to determine the polarity of Chinese text. Xu et al.[3] proposed Transfer Ada-boost and Transfer Self-training algorithms for improving the CLOA performance through selectively using the high quality transferred samples and filtering the low quality transferred samples, respectively. Furthermore, Dai [4] adopted Boosting technology to improve the system performance on target language through enhancing the weights of target language samples and reducing the weight of low quality transferred source language samples, iteratively. A typical co-training based CLOA works is conducted by Wan [5, 6], which used straight push co-training learning method for improving the opinion analysis on both languages using bilingual corpora. J. Wiebe [7] extended the bilingual learning to multi-lingual (Arabic, German, French, etc.) for improving multiple monolingual opinion analysis following co-training strategy.

Current CLOA research has yielded obvious progress, but there are still many problems left. Attribute to the small annotated opinion corpus on target language, the transfer learning based CLOA sometimes falls into an over-fitting to the transferred source language samples. Meanwhile, the co-training based CLOA, which estimates the bilingual model similarity to the samples, always leads to a relatively limited performance improvement. Target to these problems, this paper proposes a mixed CLOA model, which incorporates bilingual transfer self-training and co-training method to improve the monolingual opinion classifier, respectively. The outputs of these two models are weighted to generate the final classification results based on their training error rates. This approach is evaluated on NLP&CC2013 CLOA bakeoff data set. The experimental results show that our proposed approach outperforms transfer learning and co-training based approach. It achieves the best performance on this data set based on our knowledge.

The rest of this paper is organized as follows. Section 2 presents our mixed cross-lingual opinion analysis model. Section 3 gives the evaluation results and discussions. Finally, Section 4 concludes this paper.

2 A Mixed Cross-Lingual Opinion Analysis Model

In this section, a mixed CLOA model is proposed which adopts a weighting strategy for combining the classification output of transfer learning CLOA method and co-training CLOA method. This mixed model aims to unite the advantages of these two CLOA methods for further improving the opinion analysis performance.

2.1 Cross-Lingual Self-training Model

Cross-lingual transfer self-training model is based on cross lingual resource transfer learning. Its main idea is to transfer opinion samples from source language to target language through translation and projection. During this process, the transfer self-training method conducts confidence estimation for choosing high confidence samples into target language training corpus, iteratively. Different from Xu's work [3] which only transfers source language resources to target language, in this work, we investigate bi-directional transfer, i.e. both transfer from source language to target language and transfer from target language to source language are considered.

In general, though the target language is lack of labeled data, there is large number of unlabeled data, which is called raw corpus. Based on the available raw corpus on target language and annotated corpus on source language, we may transfer raw corpus on target language to source and transfer annotated corpus on source language to target language, respectively, through machine translation. Now, we have annotated corpus and raw corpus on both languages. The self-training algorithm is then applied to source language and target language, respectively. It is an iterative training process. In each pass, the classifier (which is trained by using annotated corpus) is applied to classify the raw samples and estimate the classification confidence. The raw samples with high confidence are moved the annotated corpus with their classification labels. The classifier is then re-trained by using the expanded annotated corpus. Such iteration terminates until the predefined condition is satisfied. The description of bilingual cross-lingual self-training is given in Algorithm 1.

More detail, for the annotated corpus on source language D_{sa}, its translation to target language is D_{ta}; for the raw corpus on target language D_{tr}, its translation to source language is D_{sr}. We train the support vector machines (SVMs) classifier on source language and target language by using D_{sa} and D_{sr}, D_{ta} and D_{tr} , respectively.

Algorithm 1. Cross-lingual transfer self-training model

```
Source language side:
Input: training sample D_sa, training examples to be transferred D_sr.
Iterations K, updated training set after i iterations T_i, number of
appended samples after every iteration k.
   T_0=D_sa
   For each i ∈ [1,K]
      1) Train SVMs classifier C^(i-1) on T_(i-1)
      2) Classify D_sr using C^(i-1)
      3) Add k classified positive and negative samples with the highest
           confidence to T_(i-1) , respectively
      4) T_i=T_(i-1)
      5) Delete the k added data from D_sr
   End for

Target language side:
Input: training sample D_ta, training examples to be transferred D_tr.
Iterations K, updated training set after i iterations T_i, number of
appended samples after every iteration k.
   T_0=D_ta
   For each i ∈ [1,K]
      1) Train SVMs classifier C^(i-1) on T_(i-1)
      2) Classify D_tr using C^(i-1)
      3) Add k classified positive and negative samples with the highest
           confidence to T_(i-1) , respectively
      4) T_i=T_(i-1)
      5) Delete the k added samples from D_tr
   End for
Output: classifier C_s which has the minimum training error among all the
iteration result of source language, classifier C_t which has the minimum
training error among all the iteration result of target language
```

On the target language side, classifier C_{ta}^{0} is trained on D_{ta}. This classifier is then used to classify the samples in the raw corpus D_{tr}. Suppose that the support vectors in C_{ta}^{0} are $v_1, v_2 \ldots v_n$, for a sample in raw corpus $d_j (d_j \in D_{tr})$:

$$W_t(d_j) = \sum_{i=1}^{n} \alpha_i Kernel(v_i, d_j) \quad (1)$$

Where, α_i is parameter in SVMs, *Kernel* is the kernel function of SVMs. The physical meaning of formula (1) is the distance between input samples and hyper-plane constructed by SVMs under this model. In this study, we regard this distance as the confidence to describe the probability that input samples are correctly classified. A value greater than 0 indicates the corresponding sample is a positive sample, otherwise, negative. The larger absolute value of confidence means the greater probability that they are correctly classified.

All of the classified samples in D_{tr} are sorted according to their confidence values. k samples with the maximum and minimum results are selected as high quality positive and negative examples, respectively. They are moved from D_{tr} to D_{ta}. In the next iteration, the classifier is re-trained on D_{ta} and C_{ta}^1 is obtained. C_{ta}^1 is then applied to classify the samples in D_{tr} and then move top k positive and negative samples D_{ta}. Such procedure repeats while the termination condition satisfied. Finally, the classifier C_t is obtained which has the smallest training error rate on the target language.

In the source language side, the same strategy is adopted to obtain the final classifier C_s which has the smallest training error rate on source language.

For a given target language sample x_t, we first generate its machine translation results x_s on source language. Based on the PAC learning theory [8], the actual error rate of the classifier has a high probability to converge to training error rate together with increasing training data. Thus, x_t is classification by,

$$y = \sum_{j=source\,\&\,target}(1 - E_j)W_j(x_j) \tag{2}$$

where, $E_j(j = source\,\&\,target)$ are the training errors on source language and target language, respectively. When y is greater than zero, x_t is classified as a positive example, otherwise, negative. The physical meaning of this weighting formula is that for two classifiers, the one with lower training error rate should have a higher weight in the final decision.

2.2 Cross-Lingual Co-training Model

The main procedure of cross-lingual co-training model is similar to cross-lingual transfer self-training model. The major difference is that transfer self-training regards the classifier training in both languages as two independent iterative processing, while in the co-training model, the classification results for a sample in one language and its translation in another language are incorporated for classification in each iteration. The model is described as Algorithm 2.

More detail, for the annotated corpus on source language D_{sa}, its translation to target language is D_{ta}; for the raw corpus on target language D_{tr}, its translation to source language is D_{sr}. The support vector machines (SVMs) classifiers, C_{sa}^0 and C_{ta}^0, are trained on annotated date in source language and target language, respectively.

For a sample in raw corpus on target language, $d_j(d_j \in D_{tr})$ and its corresponding translated samples in source language $d'_j(d'_j \in D_{sr})$, their classification may be determined by following formula:

$$y = W_s(d'_j) + W_t(d_j) \tag{3}$$

Where, W_s and W_t are the weights corresponding to source language side and target language side, respectively. The outputs for each d_j are sorted according to their values of y. The top k samples with the maximum and minimum y values and their corresponding translated samples are moved from raw corpus to annotated corpus on both languages, respectively. The classifier on target language is retrained on the annotated corpus and C_{ta}^1 is obtained. Similarly, the classifier on source language is trained and $C_{sm}^1$is obtained. Such training process is repeated until the terminate condition satisfies. Finally, the classifiers, C_t and C_s, for target language and source language are obtained, respectively.

Algorithm 2. Cross-lingual co-training

Input: Source language training samples D_{sa}, training examples to be transferred D_{sr}, target language training sample D_{ta}, training examples to be transferred D_{tr}. Iterations K, updated source language training set after i iterations T_i^s, updated target language training set after i iterations T_i^t, number of appended samples after every iteration k.

$T_0^s = D_{sa}$, $T_0^t = D_{ta}$

For each $i \in [1, K]$

1) Train SVMs classifier C_{i-1}^s on T_{i-1}^s
2) Train SVMs classifier C_{i-1}^t on T_{i-1}^t
3) Classify D_{sr} using C_{i-1}^s
4) Classify D_{tr} using C_{i-1}^t
6) Add k classified positive and negative samples with the highest confidence to T_{i-1}^t and T_{i-1}^s, respectively
5) $T_i^s = T_{i-1}^s$, $T_i^t = T_{i-1}^t$
6) Delete the k added data from D_{sr} and D_{tr}

End for

Output: classifier C_s which has the minimum training error among all the iteration result of source language, classifier C_t which has the minimum training error among all the iteration result of target language

---

For the target language samples to be classified x_t, its translation to source language $x_t{}'$is generated. Its classification is then determined by following formula (2).

2.3 Weighting Strategy in Mixed Model

Considering that the transfer self-training based model regards target language and source language as two separated languages, sometimes it falls to over fitting the training error in single language side. On the other hand, co-training based model obtained the improvements on both source and target languages. However, the achieved classification performance improvement by following formula (2) is always lower than transfer training model. Therefore, we propose a mixed model with weighing strategy in order to incorporate the classification results by transfer self-training based model and co-training based model.

The weighting strategy is similar to formula (2). Based on the PAC learning theory [8], we know that, with the increasing training samples, the actual classifier error rate will converge to the training error rate with a high probability:

$$Pr\left(|error_{train} - error_{test}| < \sqrt{\frac{ln\left(\frac{1}{\delta}\right)}{2m}}\right) \geq 1 - \delta \tag{4}$$

Where m is the number of training samples, δ is any positive number.

In this study, transfer self-learning model and co-training model have the same sample amounts, thus the actual error rates of the two models will have the same probability to converge to their own training error rates. Therefore, the error rates are used as weighting parameters for classifications results of two models. The model which has lower training error rate is assigned a higher weight in the voting.

Assume that after k iterations of transfer self-training, the trained classifier with the minimum training error in k trained classifiers on source language is labeled as $C_{st:s}^{k}$ and the trained classifiers with the minimum training error in k trained classifiers on target language is labeled as $C_{ct:t}^{k}$. Similarly, after k iterations of co-training, the trained classifiers with the minimum training error in k trained classifiers on source language is labeled as $C_{ct:s}^{k}$ and the trained classifiers the minimum training error in k trained classifiers on target language is labeled as $C_{ct:t}^{k}$. For a target language sample to be classified, x_t and its translations x_s in source language, the results of function W corresponding to the above four classifiers are calculated by following formula (1). The final classification is determined by weighting the classification results of the four classifiers as given in formula (5):

$$y = \sum_{j=s\&t} \sum_{i=st\&ct} (1 - E_{i:j}) W_{i:j}^{k}(x_j) \tag{5}$$

Similar to formula (2), when y is greater than zero, the sample sentence will be classified as positive, otherwise negative. In formula (5), the classifier with a lower error rate is assigned a higher weight in the final voting because it leads to higher probability to classification with lower error rate. This mixed method takes the confidence and training error rate into account is expected to combine multiple classifier outputs for a better performance.

3 Experiment Results and Analysis

3.1 Experiment Settings

The proposed mixed CLOA model is evaluated on NLP&CC 2013 CLOA bakeoff dataset. This dataset consists of the reviews on DVD, Book and Music category. The training data of each category contains 4,000 English annotated documents (ratio of positive and negative samples is 1:1) and Chinese raw corpus contains 17,814 DVD documents, 47,071 Book documents and 29,677 Music documents. In the testing dataset, each category contains 4,000 Chinese documents. The performance is evaluated by the correct classification accuracy for each category, and the average accuracy of the three categories, respectively.

The category accuracy is defined as:

$$Accuracy_c = \frac{\#correctly\ classified\ samples\ in\ category\ c}{4000} \tag{6}$$

Where c represent one of the DVD, Book and Music categories, respectively. The overall average accuracy is defined as:

$$Accuracy = \frac{1}{3}\sum_c Accuracy_c \tag{7}$$

In this experiment, ICTCLAS is used as the word segmentation tool. The monolingual opinion classifiers are developed based on SVMs (using SVM^light[1]) while word unigram and word bigram features are employed.

Firstly, we directly use the translated source language annotated data as the training examples. The achieved baseline performances listed in Table 1:

Table 1. Baseline performance (directly using translation results)

Category	Accuracy
$\text{Accuracy}_{\text{DVD}}$	0.7373
$\text{Accuracy}_{\text{Book}}$	0.7215
$\text{Accuracy}_{\text{Music}}$	0.7423
Accuracy	0.7337

The performances of transfer self-training model, co-training model and mixed model are then evaluated and discussed in the following subsections, respectively. In the experiment, the maximum number of iterations is set to 200. The numbers of added positive and negative samples in each iteration is set to10.

3.2 Evaluation on Transfer Self-training Model

In Experiment 1, the performance of transfer self-training model is evaluated. The achieved performances corresponding to iteratively trained classifiers are shown in Figure 1.

It is observed that the performances of transfer self-training model improved with the increasing of training iterations on both Chinese and English classifiers, but the overall performance grows slowly. Since the two classifiers on different languages

[1] `http://svmlight.joachims.org/`

are regarded independent in transfer self-training model, the weighted voting results by these two monolingual classifiers may obtain better result. In DVD category, the weighted classifier voting achieves the 2.4% and 4.6% further accuracy improvements on Chinese and English classifier, respectively. In Book category, the classifier voting obtained 5% and 3.4% further improvements, respectively. In Music category, the classifier voting obtained a lower advantage that only 0.2% higher than the Chinese classifier, but 6.3% higher than English classifier.

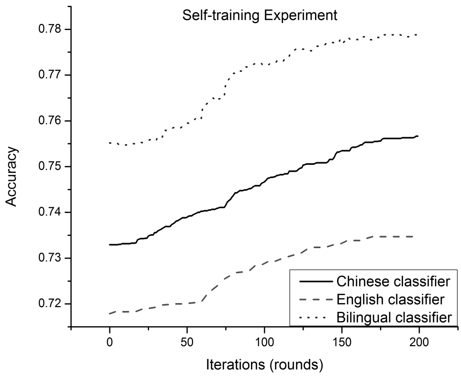

Fig. 1. Performance of transfer self-training model

Overall, the final performance of transfer self-training model achieves 2-4.7% improvement on monolingual classifier.

3.3 Evaluation of Co-training Model

Experiment 2 evaluates the co-training model. The achieved performances corresponding to iteratively trained classifiers are shown in Figure 2. In DVD category, the accuracy of Chinese classifier increased by 4.6% and the accuracy of English classifier increased by 4.7%. In Book category, the accuracy improvement for Chinese and English are 5.8% and 3.3%, respectively. In Music category, the accuracy improvements are 3.4% and 6.3%, respectively. In general, co-training model leads to obvious classifier performance improvement. The average accuracy increased by 4.7%, which is much higher than transfer self-training model. The main reason is that, during the co-training process of the two classifiers on the two languages, the selected transfer samples are more reliable by considering the confidence from both classifiers. Thus, the performance improvement of the final model is obvious.

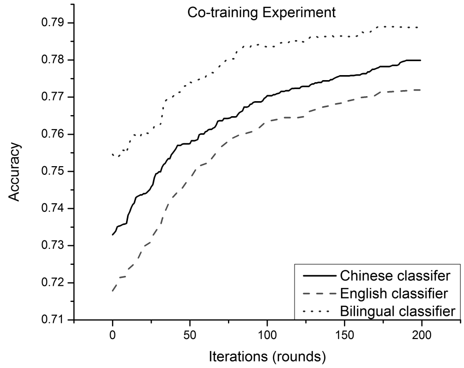

Fig. 2. Performance of co-training model

However, when applying weighted voting strategy to monolingual classifiers on the two languages, the performance improvement is limited. In DVD category, the result by weighted voting is only 1.6% higher than Chinese monolingual classifier and 1.8% higher than English monolingual classifier, respectively. In the Book category, the accuracy improvements are 1.2% and 1.3%, respectively. In Music category, the accuracy improves 0.1% and 2.5%, respectively.

In general, the co-training model leads to 1-1.7% accuracy improvement. Compared to transfer self-training model, the further performance improvement by co-training model is limited.

3.4 Evaluation on Mixed CLOA Model

Experiment 3 evaluates the mixed CLOA model which incorporates transfer self-training model and co-training model. The achieved performances corresponding to iteratively trained models are shown in Figure 3.

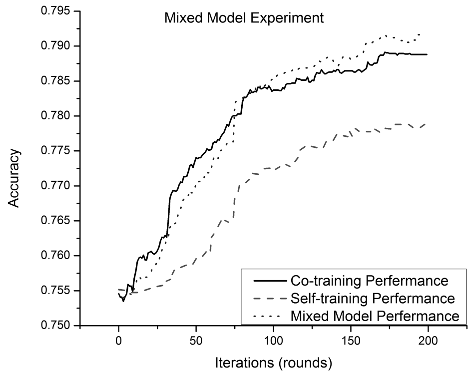

Fig. 3. Performance of the mixed CLOA model

It is observed that generally the mixed model further improve the classifier performances. In DVD category and Music category, the mixed model performance improves about 0.2% compared to the best single model classifier. In the book category, the mixed model achieved 0.6% further improvement. The reason for these different performance improvements results is as follows. In the weighted voting framework, the model which performs a better performance is assigned a higher weight. Thus, for general samples classification, the better single model plays a major role in voting. When the better single model outputs a low confidence and another model has a higher confidence, the second model plays the major role. Hence, the mixed voting strategy is shown effective to avoid the risks caused by low performance of single model.

3.5 Compared with Other Results on NLP&CC 2013 CLOA Dataset

Six teams participate the NLP&CC 2013 cross lingual opinion analysis bakeoff. The achieved performance of each team are listed in Table 2 and compared with our proposed CLOA model.

In NLP&CC 2013 CLOA bakeoff, HLT-Hitsz achieved the best accuracy performance. This system is developed by our team in the bakeoff. Meanwhile, it is shown that our proposed mixed CLOA model further improves the performance. The achieved accuracies are higher than the listed system in all of the three categories. Up to now, the final overall performance is the best result on this dataset base on our knowledge.

Table 2. Performance comparision on NLP&CC2013 CLOA dataset

Team	DVD	Music	Book	Accuracy
BISTU	0.6473	0.6605	0.5980	0.6353
HLT-Hitsz	0.7773	0.7513	0.7850	0.7712
THUIR-SENTI	0.7390	0.7325	0.7423	0.7379
SJTUGSLIU	0.7720	0.7453	0.7240	0.7471
LEO_WHU	0.7833	0.7595	0.7700	0.7709
Our Approach	**0.7965**	**0.7830**	**0.7870**	**0.7889**

4 Conclusion

This paper proposes a mixed cross-lingual opinion analysis model which weighted incorporates transfer self-training model and co-training model. This mixed model achieves the best performance on NLP&CC 2013 CLOA bakeoff dataset which shows the effectiveness of our proposed mixed CLOA model.

Since the transfer learning process does not satisfy the independent identical distribution hypothesis of training samples and test samples, actually our proposed weighted strategy based on training error rate is a kind of estimation of theory weighting. Meanwhile, the further performance improvement by following this weighting strategy is limited which is shown in the experiments. Therefore, the strategy for accurately filtering samples and estimating classifier error rate are the important problems to be solved in our future study.

References

1. Mihalcea, R., Banea, C., Wiebe, J.: Learning Multilingual Subjective Language via Cross-lingual Projections. In: Proceedings of the Association for Computational Linguistics 2007, pp. 976–983 (2007)
2. Yao, J.X., Wu, G.F., Liu, J., et al.: Using Bilingual Lexicon to Judge Sentiment Orientation of Chinese Words. In: Proceedings of the 6th International Conference on Computer and Information Technology, pp. 38–43 (2006)
3. Xu, J., Xu, R.F., Ding, Y.X., et al.: Cross Lingual Opinion Analysis via Transfer Learning. Australian Journal of Intelligent Information Processing Systems 11(2), 28–34 (2010)
4. Dai, W.Y., Yang, Q., Xue, G.R., et al.: Boosting for Transfer Learning. In: Proceedings of the 24th International Conference on Machine Learning, pp. 193–200 (2007)
5. Zhou, X.J., Wan, X.J., Xiao, J.G.: Cross-Language Opinion Target Extraction in Review Texts. In: Proceedings of IEEE 12th International Conference on Data Mining, pp. 1200–1205 (2012)
6. Wan, X.J.: A Comparative Study of Cross-Lingual Sentiment Classification. In: Proceedings of Web Intelligence, pp. 24–31 (2012)
7. Mihalcea, R., Banea, C., Wiebe, J., et al.: Multilingual Subjectivity Analysis Using Machine Translation. In: Proceedings of the 18th ACM Conference on Information and Knowledge management, pp. 127–136 (2009)
8. Valiant, L.: A Theory of the Learnable. Communications of the ACM 27(11), 1134–1142 (1984)