A POS-based Ensemble Model for Cross-domain Sentiment Classification

Rui Xia and **Chengqing Zong**

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
rxia.cn@gmail.com, cqzong@nlpr.ia.ac.cn

Abstract

In this paper, we focus on the tasks of cross-domain sentiment classification. We find across different domains, features with some types of part-of-speech (POS) tags are domain-dependent, while some others are domain-free. Based on this finding, we proposed a POS-based ensemble model to efficiently integrate features with different types of POS tags to improve the classification performance. Weights are trained by stochastic gradient descent (SGD) to optimize the perceptron and minimal classification error (MCE) criteria. Experimental results show that the proposed ensemble model is quite effective for the task of cross-domain sentiment classification.

1 Introduction

In recent years, transfer learning and domain adaptation, the task aiming to utilize labeled data from the other domains (source domain) to help learning for current domain (target domain), has attracted more and more attention in the fields of both machine learning and natural language processing, including sentiment classification. The task of sentiment classification is supposed to be domain-specific. Classifiers trained on the source domain usually perform poorly in the target domain. This is quite reasonable since the word distribution changes from one domain to another, and some words that are positive in one domain may express an opposite meaning in another one. Therefore, it is challenging to transfer a classifier trained on the source domain to the target domain.

Methodology to solve this problem can be divided into three major categories (Pan and Yang, 2009): the instance-based transfer, the feature-based transfer and the parameter-based transfer. The instance-based transfer learns the importance of labeled data in the source domain by instance re-weighting and importance sampling. These re-weighed instances are then used for learning in the target domain. Feature-based transfer aims to learn a good feature representation for the target domain using labeled data in the source domain with the help of a large number of unlabeled data in the target domain. The parameter-based transfer mostly assumes that individual models for related tasks share some parameters or prior distribution of hyper-parameters. The shared part is then added to the cost function for transfer learning.

In this paper, we propose a POS-based ensemble model for cross-domain sentiment classification. Other than the above-mentioned methodology, the transfer procedure in our approach is neither instance re-weighting nor feature representation. Broadly speaking, our approach belongs to the parameter-based transfer, but different from the traditional ways, the transfer procedure is embodied in an ensemble manner.

By observing the K-L distance of multi-domain datasets, we find that cross different domains, the distribution of features with some types of POS tags, such as adjectives and adverbs, has little

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 614–622,
Chiang Mai, Thailand, November 8 – 13, 2011.

change; while some other parts, for example, nouns, vary sharply. Furthermore, we investigate the most significant features ranked by information gain (IG). We find that the significance of adjectives and adverbs increases from in-domain to cross-domain tasks, while nouns become less important.

Based on these findings, we infer that an efficient ensemble of features according to their POS tags, may benefit more from the domain-free parts and overcome the drawbacks of domain-dependent parts, and finally enhance the overall cross-domain sentiment classification performance. We proposed two methods, namely the average perceptron (Perc) and the minimal classification error (MCE) criterion, to learn the weights of base-classifiers.

The remainder of this paper is organized as follows. Section 2 reviews related work. In Section 3, we introduce our motivation with detailed investigation. In Section 4, we propose our ensemble model for cross-domain sentiment classification. Experimental results are reported and discussed in Section 5 and 6 respectively. Section 7 draws conclusions and outlines directions for future work.

2 Related Work

2.1 Domain Adaptation

Existing approaches for cross-domain sentiment classification mostly belong to the feature-based transfer. Among them, the structural correspondence learning (SCL) algorithm proposed by (Blitzer et al., 2007) is the representative one. SCL tries to get the mapping matrix from non-pivot feature space to pivot feature space. Non-pivot features are then transferred though a projection over the principle components of the mapping matrix. (Li et al., 2009b) proposed to transfer lexical prior knowledge across domains via matrix factorization techniques. (Pan et al., 2010) proposed cross-domain sentiment classification via spectral feature alignment and compared their method with SCL.

Another work (Aue and Gamon, 2005) combined small amounts of labeled data with large amounts of labeled data in target domain to learn the model parameters for a generative naïve Bayes classifier using the Expectation Maximization (EM) algorithm.

The above work all need a large amount of unlabeled data in the target domain to help building the transfer procedure. Our approach does not need those unlabeled data. Nevertheless, we do need a small amount of labeled data from target domain, say, 50-200 instances, to help transfer learning.

2.2 Ensemble Techniques

Several researchers have achieved improvements in sentiment classification accuracy via the ensemble techniques. The work (Whitehead and Yaeger, 2008) conducted four ensemble algorithms (bagging, boosting, random subspace and bagging random subspaces) for sentiment classification. In the work by (Li et al., 2007), different classifiers were generated with different sets of features according to their POS tags. Those component classifiers are then selected and combined using several fixed rules. Experimental results showed that sum rule achieves the best performance.

We made a comparative study (Xia et al., 2011) about the effectiveness of ensemble techniques for sentiment classification. Two schemes of feature set were designed at first. Three well-known classification algorithms were then employed as base-classifiers for each of the feature sets. Three types of ensemble models were finally conducted for three ensemble strategies, with the emphasis on the evaluation of the effectiveness of ensemble techniques for sentiment classification.

Different from above methods, our focus in this paper is cross-domain sentiment classification. Compared to our former reports, we prove that the ensemble model is more effective for cross-domain tasks than for the in-domain ones.

3 Problem Investigation

3.1 POS Tag Groups

The POS information is supposed to be a significant indicator of sentiment expression. The work on subjectivity detection (Hatzivassiloglou and Wiebe, 2000) revealed a high correlation between the presence of adjectives and sentence subjectivity, yet this may not be taken to mean that other POS tags do not contribute. Indeed, it was resulted in (Pang et al., 2002; Benamara et al., 2007) that using only adjectives as features actually results in much worse performance than using the same number of most frequent unigrams. Other re-

searchers (Riloff et al., 2003) pointed out that certain verbs and nouns are also strong indicators of sentiment. According to their significance to sentiment classification, we categorize the POS tags into four groups, as shown in Table 1.

Group	Contained POS tags
J	adjectives, adverbs
V	verbs
N	nouns
O	the other POS tags

Table 1. Four groups of POS tags

3.2 Cross-domain K-L Distances

When conducting transfer learning, it is crucial to find that from one domain to another, which part of knowledge changes and which part of knowledge remains similar. Then the "unchanged" part of knowledge should be kept during the learning process, while the "changed" part should be transferred. Our intuition is that from one domain to another, nouns change the most, because domains (or topics) are mostly denoted by nouns; while adjectives and adverbs change less, for example, "*great*" and "*love*" always express the meaning that something is good, no matter the domain is Book or Movie.

Holding this belief, we observe the cross-domain K-L distance (also called relative entropy) of the class-conditional distribution of each type of POS tags. We use the Multi-Domain Sentiment Dataset[1] for statistics. This dataset was introduced by (Blitzer et al., 2007) and then widely used in the field of cross-domain sentiment classification. It contains product reviews taken from Amazon.com from four product types (domains) – Book (B), DVD (D), Electronics (E) and Kitchen (K). Each of these contains 1000 positive and 1000 negative reviews.

We use the term "X-Y" to denote the task computing K-L distance of domain X and Y. For example, "B-D" denotes the K-L distance between the Book domain and DVD domain. We compute the K-L distance of two domains for each class based on the assumption that the class-conditional distribution is the multinomial distribution. The results are presented in Table 2.

We focus on the comparison between different types of POS tags. The K-L distance of N is the largest in all cross-domain tasks, significantly larger than the other POS types and Uni (unigrams). It indicates that from one domain to another, the change of N is the biggest part. On the contrary, the distribution of O changes the least. It is reasonable that the POS tags contained in O, such as prepositions, pronouns, etc., are mostly domain-free. The K-L distance of J is larger than that of O, but significantly smaller than that of N. The value is also smaller compared to that of all unigrams. V gives the comparable K-L distance. We may conclude that most features in J and V are partially domain-free. It also coincides with our intuition that "*great*" and "*love*" always express a positive meaning in whatever domains.

Generally, the cross-domain K-L distances of different types of POS tags can be ranked as: N>>Uni>V>J>O.

Task	Class	J	V	N	O	Uni
B-D	**Pos**	0.1608	0.2022	0.5420	0.0197	0.1968
	Neg	0.1427	0.1632	0.5149	0.0144	0.1779
B-E	**Pos**	0.4353	0.3752	1.2125	0.1329	0.4738
	Neg	0.3585	0.3414	1.1787	0.1221	0.4416
B-K	**Pos**	0.4487	0.4255	1.2059	0.1146	0.4752
	Neg	0.3348	0.3770	1.2620	0.1298	0.4690
D-E	**Pos**	0.3983	0.3614	1.1751	0.1281	0.4579
	Neg	0.3430	0.3429	1.1850	0.0905	0.4279
D-K	**Pos**	0.4028	0.4125	1.2587	0.1168	0.4820
	Neg	0.3372	0.3687	1.3352	0.0921	0.4686
E-K	**Pos**	0.2428	0.1934	0.9310	0.0208	0.3093
	Neg	0.1856	0.1836	0.7791	0.0153	0.2592

Table 2: Cross-domain K-L distance

3.3 Most Significant Cross-domain Features

Furthermore, we investigate the most significant cross-domain features. We choose the top-N features that are ranked by information gain (IG) which was proved to be an effective feature selection method for sentiment classification (Li et al.,

[1] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

2009a). In table 3, we report the number of different POS tags from top-50, 100, 200, 500 and 1000 features respectively. "In" denotes the average result of four individual domains. "Share" denotes the number of features shared by all of the four groups of top-*N* features.

We first observe the average results of four individual top-100 features. The number of J, V, N and O cover the percentage of 42.0, 24.0, 24.0 and 9.0 respectively. Among the four groups of top-100 features, only 11 words appear in all of them. These features are "*great*", "*love*", "*unfortunately*", "*money*", "*highly*", "*bad*", "*worst*", "*excellent*", "*not*", "*waste*" and "*best*", where adjectives, verbs and nouns cover 81.8%, 9.1% and 9.1% respectively. In the case of individual top-200 features, the number of shared words by four domains is 19, 63.2% of which are adjectives.

As *N* increases, the percentage of four groups of POS tags in shared features can be generally ranked as: J>V>N>O. This has confirmed our intuition that nouns are the most domain-specific, while adjectives and adverbs are especially good cross-domain features.

Top-*N*	In/Share	Num	J (%)	V (%)	N (%)	O (%)
100	**In**	100	42.0	24.0	24.0	9.0
	Share	11	81.8	9.1	9.1	0.0
200	**In**	200	35.0	27.5	28.0	19.5
	Share	19	63.2	15.8	10.5	10.5
500	**In**	500	30.4	26.6	33.6	9.4
	Share	35	54.3	20.0	14.3	11.4
1000	**In**	1000	27.4	26.8	37.8	8.0
	Share	67	44.8	22.4	20.9	11.9

Table 3: Top-features by feature selection

4 The Ensemble Model

4.1 A POS-based Weighted Combination

The pursuit of POS-based weighted combination is motivated by the intuition that an appropriate integration of different participants might leverage distinct strengths. For example, the weights assigned to adjectives and adverbs are supposed to be higher than that of nouns.

We first build a new meta-feature vector $\hat{\mathbf{x}} = [o_{11},\ldots,o_{kj},\ldots,o_{DC}]$, where $o_{kj}(\mathbf{x})$ denotes the predicted score of the *k*th base-classifier for the *j*th class, C is the number of classes and D is the number of base-classifiers (in our approach C equals 2 and D equals 4). Then, the weighted combination could be represented by

$$O_j = \sum_{k=1}^{D} \omega_k o_{kj} = \sum_{k=1}^{D} \omega_k \hat{\mathbf{x}}_{k\times D+j}, \qquad (1)$$

where $\hat{\mathbf{x}}_{k\times D+j}$ denotes the score for the *j*th class of the *k*th base-classifier.

4.2 Weight Optimization

To learn the weights in Equation (1) , we propose to use stochastic gradient descent (SGD) to optimize some criteria. We consider two criteria in our approach, namely the perceptron (Perc) model, and minimal classification error (MCE) criterion.

The cost function of Perc in multi-class case is given by

$$J_p = \frac{1}{N}\sum_{i=1}^{N}\left[\max_{j=1,\ldots,C} g_j(\hat{\mathbf{x}}_i) - g_{y_i}(\hat{\mathbf{x}}_i)\right]. \qquad (2)$$

Note that in implementation, we utilize the average perceptron, a variation of perceptron that averages weights of all iterations, to improve the robustness.

The MCE criterion proposed by (Juang and Katagiri, 1992) is supposed to be more relevant to the classification error. In their approach, a simple version of misclassification measure of the instance $\hat{\mathbf{x}}_i$ from the *j*th class is defined by

$$d_j(\hat{\mathbf{x}}_i) = -g_{y_i}(\hat{\mathbf{x}}_i) + \max_{h\neq j}\{g_h(\hat{\mathbf{x}}_i)\}. \qquad (3)$$

Based on this measure, the cost function of MCE is given by

$$J_{mce} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} I(y_i = j)\delta(d_j(\hat{\mathbf{x}}_i) + \alpha), \qquad (4)$$

where $\delta(\bullet)$ is the sigmoid function, and α is the hyper-parameter.

SGD uses approximate gradients estimated from subsets of the training data and updates the parameters in an online manner:

$$w_k(t+1) = w_k(t) - \eta(t)\frac{\partial J}{\partial w_k}, \quad (5)$$

where t denotes the iteration step and $\eta(t)$ denotes the learning rate. Compared to standard gradient descent, SGD is much faster and more efficient, especially for large datasets.

5 Experiments

5.1 Experimental Settings

We use the Multi-domain dataset for experiments, which was already introduced in Section 3.2. The term "source→target" is used to denote the cross-domain tasks. For example, "D→B" represents the task that is trained in the DVD domain but the tested in the Book domain.

In our experiments, each dataset is split into a training set of 1600 instances, and a test set of 400 instances. The NLTK toolkit[2] is used for word tokenization. The MXPOST[3] tool is chosen as our POS tagger. Features with the term frequency no less than four are selected for classification.

Since it was reported that Naïve Bayes performs the best among three classifiers (Naïve Bayes, MaxEnt and SVM) on Multi-Domain Sentiment Dataset (Xia et al., 2011), we choose it as the base classification algorithm. We use the tool OpenPR-NB[4] in our experiments, with the settings of multinomial event model and Laplace smoothing.

After base-classification, the predicted score of the test set is randomly split to a meta-development set and meta-test set. The ensemble systems are trained on the meta-development set to classify the meta-test set to get the final prediction. The learning rate of SGD is set to be one and the maximal iteration number is set to be 100.

The process of meta-learning and test is randomly repeated for 100 times. All of the following results are in terms of an average of the 100 repeats.[5]

[2] http://www.nltk.org/
[3] http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html
[4] http://www.openpr.org.cn/
[5] The leave-one-out cross validation procedure was used in some previous work. In our experiments, since the size of development set is required to be comparatively small, cross validation is not quite suitable.

5.2 Results of Uni-based Ensemble

Table 4 reports the performance of Uni-based ensemble. Unigrams are categorized into four groups according to Table 1, denoted by Uni-J, Uni-V, Uni-N and Uni-O respectively. The perforamce of using all unigrams without transfer (denoted by Uni) is taken as the baseline. In ensemble approaches, we report results of three rules, i.e., the Sum rule, Perc and MCE criteria. Perc and MCE are trained by 200 labeled data in the target domain.

At first, we focus on the comparison of Uni and Uni-J. Uni-J performs consistently better than Uni in most of the cross-domain tasks (71.40% vs. 70.54%). This is opposite to the conclusion in in-domain tasks that using only adjectives as features results in much worse performance than using the same number of most frequent unigrams (Pang et al., 2002; Benamara et al., 2007). It also confirms our intuition that adjectives and adverbs are more effective feature for cross-domain tasks, and an efficient ensemble of these POS tags may be more effective.

Secondly, we observe the performance of Sum rule. We have drawn the conclusion in (Xia et al., 2011) that Sum rule is a low-cost yet effective approach for sentiment classification. However, this conclusion may not hold in the cross-domain tasks. Sum rule performs significantly worse than the best base-classifier (Uni-J). This is quite reasonable that assigning equal weights to unbalanced component base-classifiers will reduce the effect of ensemble.

Finally, we observe the results of weighted combination. Their performance is consistently higher than Uni and Uni-J, except for the task E→K (a slight decline). In average of 12 tasks, Perc and MCE outperform the Uni baseline by 3.01% and 3.94% respectively. Comparing Perc and MCE, the performance of MCE is more attractive, 0.93% higher than Perc in average.

Tasks	Uni-J	Uni-V	Uni-N	Uni-O	Uni	Ensemble		
						Sum	Perc	MCE
D→B	75.50	62.00	62.00	56.00	73.25	74.75	77.87	**78.88**
E→B	72.75	59.50	59.75	57.00	69.75	71.25	**72.35**	**72.35**
K→B	68.75	59.50	57.75	54.50	67.75	66.75	69.59	**70.47**
B→D	74.75	64.00	65.00	66.75	74.50	75.00	77.46	**77.81**
E→D	70.25	57.25	52.50	56.50	67.50	66.75	71.47	**72.66**
K→D	71.25	60.25	58.50	56.75	73.75	73.50	76.28	**77.25**
B→E	67.50	56.50	53.00	55.50	63.25	63.75	68.36	**69.26**
D→E	66.00	56.75	58.50	48.50	61.75	63.50	67.21	**68.71**
K→E	77.50	67.00	63.25	60.25	74.75	75.00	77.79	**79.74**
B→K	69.50	60.50	60.75	55.00	69.00	70.50	70.14	**71.14**
D→K	67.75	62.25	61.00	52.75	71.00	70.25	74.67	**75.86**
E→K	75.25	68.25	67.50	57.00	**80.25**	77.50	79.39	79.65
Average	71.40	61.15	59.96	56.38	70.54	70.71	73.55	**74.48**

Table 4. Performance (%) of Uni-based Ensemble

5.3 Results of UB-based Ensemble

We still consider using unigrams and bigrams together as candidate features for ensemble. Unigrams and bigrams are divided into four subsets (UB-J, UB-V, UB-N and UB-O), according to the POS tags of its headword. The performance of unigrams and bigrams without transfer is used as the baseline (denoted by UB). The reported ensemble results are also with the help of 200 labeled data from the target domain. Detailed results are presented in Table 5.

We still first compare UB-J and UB. This time, UB-J beats UB in some tasks, but it does not show general superiority. It is probably due to that some adjective information has coupled with other POS tags, such as J-N. Nevertheless, its performance is still comparative higher compared with the other three types of POS tags.

With regard to the ensemble methods, the performance of sum rule is not so sound, the same as before. The weighted combination still gains significant improvements over the UB baseline. In average, Prec and MCE outperform the UB baseline by 3.01% and 3.94% respectively. Overall, the ensemble model is quite effective for cross-domain sentiment classification. Among them, MCE is the most effective.

6 Discussion

6.1 Ensemble Model Revisited

In this section, we try to give some explanations about why the ensemble model is effective for cross-domain sentiment classification.

In traditional linear classifiers, the weights assigned to each feature are trained on the source-domain labeled data. Each weight thus embodies the significance of its responding feature to the source-domain classification. Transferring from one domain to another, those weights need to be adapted to the target domain.

Based on the observation that some parts of the feature are domain-dependent and some parts are domain-free, an efficient ensemble may be an effective way to adapt those weights to the target-domain. The behind transferring procedure can be interpreted as:

$$\begin{aligned} &\log P(c_j \mid d) \\ &\propto \log P(c_j) + \sum_i \log P(t_i \mid c_j) \\ &= \log P(c_j) + \omega_1 \sum_{i_1 \in J} \log P(t_{i_1} \mid c_j) + \omega_2 \sum_{i_2 \in V} \log P(t_{i_2} \mid c_j) \\ &\quad + \omega_3 \sum_{i_3 \in N} \log P(t_{i_3} \mid c_j) + \omega_4 \sum_{i_4 \in O} \log P(t_{i_4} \mid c_j), \end{aligned} \tag{6}$$

Tasks	UB-J	UB-V	UB-N	UB-O	UB	Ensemble		
						Sum	Perc	MCE
D→B	78.00	64.25	65.75	59.00	76.25	77.00	79.51	**81.01**
E→B	72.50	64.75	64.75	62.75	77.75	76.50	77.09	**77.80**
K→B	70.25	61.00	65.00	55.75	72.25	72.25	73.37	**74.06**
B→D	75.00	70.75	67.25	65.50	77.00	78.25	78.21	**79.03**
E→D	71.00	62.75	58.00	55.75	73.25	72.75	73.98	**74.77**
K→D	72.25	59.00	60.75	61.00	73.00	76.00	75.66	**77.14**
B→E	67.25	62.50	58.00	59.00	68.50	69.00	70.78	**72.27**
D→E	69.00	60.75	58.25	49.50	65.50	66.50	69.29	**70.34**
K→E	77.25	73.50	69.00	61.50	81.25	80.25	81.71	**82.87**
B→K	72.50	63.50	62.50	58.25	74.50	74.75	73.87	**75.02**
D→K	70.00	67.75	60.50	53.75	74.75	74.75	77.37	**78.68**
E→K	78.50	70.50	67.75	57.75	79.75	79.50	80.34	**81.13**
Average	72.79	65.08	63.13	58.29	74.48	74.79	75.93	**77.01**

Table 5. Performance (%) of UB-based Ensemble

where the conditional word probability is transferred from $P(t_i \mid c_j)$ to $P(t_i \mid c_j)^{\omega_j}$, which encodes information of both the sentiment significance and cross-domain ability.

In table 6, we present the average weights trained by MCE across all tasks. We can see that the weight of J is the largest in four parts, generally a half percentage. Thereby, the conditional probability of features in J will get a comparatively larger value. Such re-assignments of parameters will be good for cross-domain tasks.

Ensemble Tasks	J	V	N	O
Uni-based	0.52	0.20	0.16	0.12
UB-based	0.45	0.21	0.16	0.18

Table 6. Average weights trained by MCE

6.2 Sensitivity on Parameter Tuning

In this section, we test the sensitivity on parameter tuning. For simplicity, we fix the weights of V and O to be 0.20 and 0.15 respectively. We use ω to denote the weight of J, and the weight of N is thus $(0.65-\omega)$. We tune the value of ω from 0 to 0.65, and observe the average accuracy of ensemble. The curve is displayed in Fig. 1.

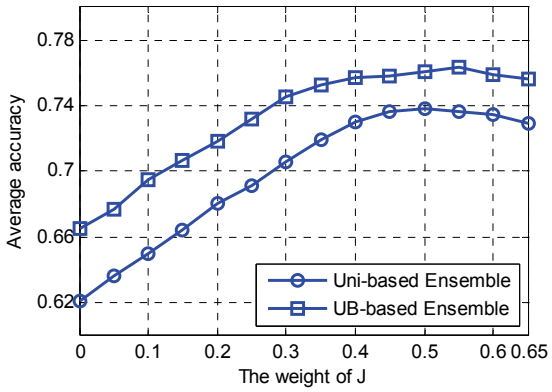

Fig. 1. Parameter sensitivity test

We can conclude from Fig. 1 that the ensemble performance is quite sensitive to the weights assigned to base-classifiers. When ω is close to 0, the weight of N is comparatively larger, and the ensemble performance drops sharply. The best result was obtained when ω locates at the area close to 0.5. The golden weights are quite similar to the results trained by MCE (Table 6). This shows that MCE is effective at parameter tuning.

Moreover, these weights can also be regarded as the empirical values when performing POS-based ensemble, in case that there is no or very few labeled data available in the target domain.

6.3 Dependency on the Size of Labeled Data in the Target Domain

Finally, we discuss the dependency of our approach on the size of labeled data in the target domain. In Fig. 2, we observe the performance of our ensemble model as the size of labeled data from the target domain increases from 50 to 300. "In-domain" denotes the accuracy trained on those labeled data for in-domain classification. "No transfer" denotes the result trained on 1600 labeled data in the source domain without transfer. Two ensemble approaches are also displayed for comparison. The reported accuracy is the average of 12 cross-domain tasks.

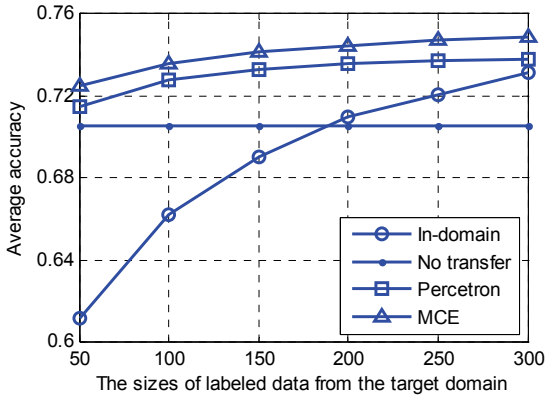

Fig. 2. Performance as labeled data in target domain increase

From Fig. 2, we can see that when the size of labeled data is small, the in-domain performance is fairly poor. At this time, the ensemble model could substantially improve the performance. As the size of labeled data increases, all of the systems yield higher performance. When the size increases to 300, the ensemble model shows limited superiority. It is reasonable that the in-domain learning is always the best if labeled data is enough.

Although the improvements of the ensemble model gained over the in-domain system become less as the size of labeled data increases, we could still conclude that in the case that there is only few labeled data in the target domain, the ensemble model is quite effective, to take the advantage of a large number of labeled data from the source domains, to help improving sentiment classification performance in the target domain.

7 Conclusion

In this paper, we propose a POS-based ensemble model for cross-domain sentiment classification. The motivation is based on the observation that some types of POS tags are domain-free, while some others are domain-dependent. Therefore, an efficient ensemble of them would leverage distinct strengths and improve the classification performance. Experimental results show that when the labeled data in the target is few, the proposed ensemble model is quite effective to make use of the labeled data from the source domain to improve the classification performance in the target domain.

We also update our previous conclusion drawn regarding the effectiveness of ensemble for in-domain sentiment classification (Xia and Zong, 2010; Xia et al., 2011). We conclude that the POS-based ensemble model is more effective for cross-domain sentiment classification than in-domain tasks.

In the future, we plan to extend the ensemble model to the tasks of cross-domain sentiment classification with multiple source domains. We also wish to make use of a large amount of unlabeled data in the target domain to help assist the ensemble performance for cross-domain sentiment classification in the framework of ensemble learning.

Acknowledgment

The research work has been funded by the Natural Science Foundation of China under Grant No. 60975053 and 61003160, and supported by the External Cooperation Program of the Chinese Academy of Sciences.

References

Anthony Aue and Michael Gamon, 2005. Customizing Sentiment Classifiers to New Domains: A Case Study. In Proceedings of Recent Advances in Natural Language Processing (RANLP).

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato and V. S. Subrahmanian, 2007. Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

John Blitzer, Mark Dredze and Fernando Pereira, 2007. Biographies, Bollywood, Boom-boxes and Blenders:

Domain Adaptation for Sentiment Classification. In Proceedings of the Association for Computational Linguistics (ACL).

Vasileios Hatzivassiloglou and Janyce Wiebe, 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 299-305.

BH Juang and S Katagiri, 1992. Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing, 40 (12). pages 3043-3054.

Shoushan Li, Rui Xia, Chengqing Zong and Chu-Ren Huang, 2009a. A framework of feature selection methods for text categorization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 692-700.

Shoushan Li, Chengqing Zong and Xia Wang, 2007. Sentiment Classification through Combining Classifiers with Multiple Feature Sets. In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pages 135-140.

Tao Li, Yi Zhang and Vikas Sindhwani, 2009b. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 244-252.

Jialin Pan, Xiaochuan Ni, Jiantao Sun, Qiang Yang and Zheng Chen, 2010. Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the International World Wide Web Conference (WWW), pages 751-760.

Jialin Pan and Qiang Yang, 2009. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22 (10). pages 1345-1359.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86.

Ellen Riloff, Janyce Wiebe and Theresa Wilson, 2003. Learning Subjective Nouns using Extraction Pattern Bootstrapping. In Proceedings of the Conference on Natural Language Learning (CoNLL), pages 25-32.

Matthew Whitehead and Larry Yaeger, 2008. Sentiment Mining Using Ensemble Classification Models. In the International Conference on Systems, Computing Sciences and Software Engineering (CISSE).

Rui Xia, Chengqing Zong and Shoushan Li, 2011. Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification. Information Sciences, 181 (6). pages 1138-1152.

Rui. Xia and Chengqing. Zong, 2010. Exploring the use of word relation features for sentiment classification. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pages 1336-1344.