Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora

Bin Lu[1,3*], **Chenhao Tan**[2], **Claire Cardie**[2] **and Benjamin K. Tsou**[3,1]

[1] Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong

[2] Department of Computer Science, Cornell University, Ithaca, NY, USA

[3] Research Centre on Linguistics and Language Information Sciences,
Hong Kong Institute of Education, Hong Kong

lubin2010@gmail.com, {chenhao, cardie}@cs.cornell.edu, btsou99@gmail.com

Abstract

Most previous work on multilingual sentiment analysis has focused on methods to adapt sentiment resources from resource-rich languages to resource-poor languages. We present a novel approach for joint bilingual sentiment classification at the sentence level that augments available labeled data in each language with unlabeled parallel data. We rely on the intuition that the sentiment labels for parallel sentences should be similar and present a model that jointly learns improved monolingual sentiment classifiers for each language. Experiments on multiple data sets show that the proposed approach (1) outperforms the monolingual baselines, significantly improving the accuracy for both languages by 3.44%-8.12%; (2) outperforms two standard approaches for leveraging unlabeled data; and (3) produces (albeit smaller) performance gains when employing pseudo-parallel data from machine translation engines.

1 Introduction

The field of sentiment analysis has quickly attracted the attention of researchers and practitioners alike (e.g. Pang et al., 2002; Turney, 2002; Hu and Liu, 2004; Wiebe et al., 2005; Breck et al., 2007; Pang and Lee, 2008). Indeed, sentiment analysis systems, which mine opinions from textual sources (e.g., news, blogs, and reviews), can be used in a wide variety of applications, including interpreting product reviews, opinion retrieval and political polling.

Not surprisingly, most methods for sentiment classification are supervised learning techniques, which require training data annotated with the appropriate sentiment labels (e.g. document-level or sentence-level positive vs. negative polarity). This data is difficult and costly to obtain, and must be acquired separately for each language under consideration.

Previous work in multilingual sentiment analysis has therefore focused on methods to adapt sentiment resources (e.g. lexicons) from resource-rich languages (typically English) to other languages, with the goal of transferring sentiment or subjectivity analysis capabilities from English to other language (e.g. Mihalcea et al. (2007); Banea et al. (2008; 2010); Wan (2008; 2009); Prettenhofer and Stein (2010)). In recent years, however, sentiment-labeled data is gradually becoming available for languages other than English (e.g. Seki et al. (2007; 2008); Nakagawa et al. (2010); Schulz et al. (2010)). In addition, there is still much room for improvement in existing monolingual (including English) sentiment classifiers, especially at the sentence level (Pang and Lee, 2008).

This paper tackles the task of bilingual sentiment analysis. In contrast to previous work, we (1) assume that some amount of sentiment-labeled data is available for the language pair under investigation, and (2) investigate methods to simultaneously improve sentiment classification for *both languages*. Given the labeled data in each language, we propose an approach that exploits an *unlabeled* parallel corpus with the following

[*] The work was conducted when the first author was visiting Cornell University.

intuition: *two sentences or documents that are parallel (i.e. translations of one another) should exhibit the same sentiment — i.e. their sentiment labels (e.g. polarity, subjectivity, intensity) should be similar*. The proposed maximum entropy-based EM approach jointly learns two monolingual sentiment classifiers by treating the sentiment labels in the unlabeled parallel text as unobserved latent variables, and maximizes the regularized joint likelihood of the language-specific labeled data together with the inferred sentiment labels of the parallel text. Although our approach should be applicable at the document-level and for additional sentiment tasks, we focus on sentence-level polarity classification in this work.

We evaluate our approach for English and Chinese on two dataset combinations (see Section 4) and find that the proposed approach outperforms the monolingual baselines (i.e. maximum entropy and SVM classifiers) as well as two alternative methods for leveraging unlabeled data (transductive SVMs (Joachims, 1999b) and co-training (Blum and Mitchell, 1998)). Accuracy is significantly improved for both languages, by 3.44%-8.12%. We furthermore find that improvements, albeit smaller, are obtained when the parallel data is replaced with a pseudo-parallel (i.e. automatically translated) corpus. To our knowledge, this is the first multilingual sentiment analysis study to focus on methods for simultaneously improving sentiment classification for a pair of languages based on unlabeled data rather than resource adaptation from one language to another.

The rest of the paper is organized as follows. Section 2 introduces related work. In Section 3, the proposed joint model is described. Sections 4 and 5, respectively, provide the experimental setup and results, followed by the conclusion in Section 6.

2 Related Work

Multilingual Sentiment Analysis. There is a growing body of work on multilingual sentiment analysis. Most approaches focus on resource adaptation from one language (usually English) to other languages with few sentiment resources. Mihalcea et al. (2007), for example, generate subjectivity analysis resources in a new language from the English sentiment resources by leveraging a bilingual dictionary or a parallel corpus. Banea et al. (2008; 2010) instead automatically translate the English resources by using automatic machine translation engines for subjectivity classification. Prettenhofer and Stein (2010) investigate cross-lingual sentiment classification from the perspective of domain adaptation based on structural correspondence learning (Blitzer et al., 2006).

Approaches that do not explicitly involve resource adaptation include Wan (2009), which uses co-training (Blum and Mitchell, 1998) with English vs. Chinese features comprising the two independent "views" to exploit unlabeled Chinese data and a labeled English corpus and thereby improves Chinese sentiment classification. Another notable approach is the work of Boyd-Graber and Resnik (2010), in which they present a generative model, supervised multilingual latent Dirichlet allocation, by jointly modeling topics that are consistent across languages, and employing them to better predict sentiment ratings.

Unlike the methods described above, we focus on simultaneously improving the performance of sentiment classification in a pair of languages by developing a model that relies on sentiment-labeled data in each language as well as unlabeled parallel text for the language pair.

Semi-supervised Learning. Another line of related work is semi-supervised learning, which combines labeled and unlabeled data to improve the performance of the task of interest (Zhu and Goldberg, 2009). Among the popular semi-supervised methods (e.g. EM on Naïve Bayes (Nigam et al., 2000), co-training (Blum and Mitchell, 1998), transductive SVMs (Joachims, 1999b), and co-regularization (Sindhwani et al., 2005; Amini et al., 2010)), our approach employs the EM algorithm, extending it to the bilingual case based on maximum entropy. We compare to co-training and transductive SVMs in Section 5.

Multilingual NLP for Other Tasks. Finally, there exists related work using bilingual resources to help other NLP tasks, such as word sense disambiguation (e.g. Ido and Itai (1994)), parsing (e.g. Burkett and Klein (2008); Zhao et al. (2009); Burkett et al. (2010)), information retrieval (Gao et al., 2009), named entity detection (Burkett et al., 2010); topic extraction (e.g. Zhang et al., 2010), text classification (e.g. Amini et al., 2010), and hyponym-relation acquisition (e.g. Oh et al., 2009).

In these cases, multilingual models increase performance because different languages contain different ambiguities and therefore present complementary views on the shared underlying labels. Our work shares a similar motivation.

3 A Joint Model with Unlabeled Parallel Text

We propose a maximum entropy-based statistical model. Maximum entropy (MaxEnt) models[1] have been widely used in many NLP tasks (Berger et al., 1996; Ratnaparkhi, 1997; Smith, 2006). It assigns the conditional probability of the label y given the observation x as follows:

$$p(y|x;\vec{\theta}) = \frac{1}{Z}\exp\left(\vec{\theta}\cdot\vec{f}(x,y)\right) \quad (1)$$

where $\vec{\theta}$ is a real-valued vector of feature weights and $\vec{f}$ is a feature function that maps pairs (x, y) to a nonnegative real-valued feature vector. Each feature has an associated parameter, θ_i, which is called its weight; and Z is the corresponding normalization factor.

Maximum likelihood parameter estimation (training) for such a model, with a set of labeled examples $\{(x_i, y_i)_{i=1}^n\}$, amounts to solving the following optimization problem:

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \prod_{i=1}^n p(y_i|x_i;\vec{\theta}) \quad (2)$$

3.1 Problem Definition

Given two languages L_1 and L_2, suppose we have two distinct (i.e. not parallel) sets of sentiment-labeled data D_1 and D_2 written in L_1 and L_2, respectively. In addition, we have unlabeled (w.r.t. sentiment) bilingual (in L_1 and L_2) parallel data U, which are defined as follows.

$$D_1 = (X_1, Y_1) = \{(x_i^1, \mathrm{y_i^1})_{i=1}^{l_1}\}$$
$$D_2 = (X_2, Y_2) = \{(x_i^2, \mathrm{y_i^2})_{i=1}^{l_2}\}$$
$$U = (X_1', X_2') = \left\{\left(x_i^{1'}, x_i^{2'}\right)_{i=1}^{u}\right\}$$

where $y_i \in \mathcal{Y} = \{+1, -1\}$ denotes the polarity of the i-th instance x_i (positive or negative); l_1 and l_2 are respectively the numbers of labeled instances in L_1 and L_2; $x_i^{1'}$ and $x_i^{2'}$ are parallel instances in L_1 and L_2, respectively (i.e. they are supposed to be translations of one another), whose labels $y_i^{1'}$ and $y_i^{2'}$ are unobserved, but according to the intuition outlined in Section 1, should be similar.

Given the input data D_1, D_2 and U, our task is to jointly learn two monolingual sentiment classifiers — one for L_1 and one for L_2. With MaxEnt, we learn from the input data:

$$f\colon \{D_1, D_2, U\} \rightarrow (\vec{\theta}_1^*, \vec{\theta}_2^*)$$

where $\vec{\theta}_1^*$ and $\vec{\theta}_2^*$ are the vectors of feature weights for L_1 and L_2, respectively (for brevity we denote them as θ_1 and θ_2 respectively in the remaining sections). In this study, we focus on sentence-level sentiment classification, i.e. each x_i is a sentence, and $x_i^{1'}$ and $x_i^{2'}$ are parallel sentences.

3.2 The Joint Model

Given the problem definition above, we now present a novel model to exploit the correspondence of parallel sentences in unlabeled bilingual text. The model maximizes the following joint likelihood with respect to θ_1 and θ_2:

$$\begin{aligned}\mathcal{L}(\theta_1, \theta_2|D_1, D_2, U) &= p(Y_1|X_1;\theta_1)p(Y_2|X_2;\theta_2)\\ &\quad p(Y_1', Y_2'|X_1', X_2';\theta_1, \theta_2)\\ &= \prod_{v=1}^2\prod_{i=1}^{l_v} p(y_i^v|x_i^v;\theta_v)\\ &\quad \prod_{i=1}^u p\left(y_i^{1'}, y_i^{2'}|x_i^{1'}, x_i^{2'};\theta_1, \theta_2\right)\end{aligned} \quad (3)$$

where $v \in \{1,2\}$ denotes L_1 or L_2; the first term on the right hand side is the likelihood of labeled data for both D_1 and D_2; and the second term is the likelihood of the unlabeled parallel data U.

If we assume that parallel sentences are perfect translations, the two sentences in each pair should have the same polarity label, which gives us:

$$\begin{aligned}&p\left(y_i^{1'}, y_i^{2'}|x_i^{1'}, x_i^{2'};\theta_1, \theta_2\right) =\\ &\textstyle\sum_{y_i'} p\left(y_i'|x_i^{1'};\theta_1\right)p\left(y_i'|x_i^{2'};\theta_2\right)\end{aligned} \quad (4)$$

where y_i' is the unobserved class label for the i-th instance in the unlabeled data. This probability directly models the agreement on sentiment labels between $x_i^{1'}$ and $x_i^{2'}$.

However, there could be considerable noise in real-world parallel data, i.e. the sentence pairs may be noisily parallel (or even comparable) instead of fully parallel (Munteanu and Marcu, 2005). In such noisy cases, the labels (positive or negative) could be different for the two monolingual sentences in a sentence pair. Although we do not know the exact probability of two sentences in each pair having the same label, we can approximate it by using

[1] They are sometimes referred to as log-linear models, but also known as exponential models, generalized linear models, or logistic regression.

their translation probabilities, which can be computed using word alignment toolkits such as Giza++ (Och and Ney, 2003) or the Berkeley word aligner (Liang et al., 2006). *The intuition here is that if the translation probability of two sentences is high, the probability that they have the same label should be high as well*. Therefore, by considering the noise in parallel data, we get:

$$p(y_i^{1'}, y_i^{2'} | x_i^{1'}, x_i^{2'}; \theta_1, \theta_2) = \sum_{y'}\{p(a_i)p(y'|x_i^{1'};\theta_1)p(y'|x_i^{2'};\theta_2)\} + \sum_{y'}\{(1-p(a_i))p(y'|x_i^{1'};\theta_1)p(\bar{y'}|x_i^{2'};\theta_2)\} \quad (5)$$

where $p(a_i)$ is the translation probability of the i-th sentence pair in U;[2] $\bar{y'}$ is the opposite of y'; the first term models the probability that $x_i^{1'}$ and $x_i^{2'}$ have the same label; and the second term models the probability that they have different labels.

By further considering the weight to ascribe to the unlabeled data vs. the labeled data (and the weight for the L2-norm regularization), we get the following regularized joint log likelihood to be maximized:

$$\log \mathcal{L}(\theta_1, \theta_2 | D_1, D_2, U) = \sum_{v=1}^{2} \log p(Y_v | X_v; \theta_v) + \lambda_1 \log p(Y_1', Y_2' | X_1', X_2'; \theta_1, \theta_2) - \frac{\lambda_2}{2} \sum_{v=1}^{2} \| \theta_v \|_2^2 \quad (6)$$

where the first term on the right hand side is the log likelihood of the labeled data from both D_1 and D_2; the second is the log likelihood of the unlabeled parallel data U, multiplied by $\lambda_1 \geq 0$, a constant that controls the contribution of the unlabeled data; and $\lambda_2 \geq 0$ is a regularization constant that penalizes model complexity or large feature weights. When λ_1 is 0, the algorithm ignores the unlabeled data and degenerates to two MaxEnt models with only the labeled data.

3.3 The EM Algorithm on MaxEnt

To solve the optimization problem for the model, we need to jointly estimate the optimal parameters for the two monolingual classifiers by finding:

$$(\theta_1^*, \theta_2^*) = \arg\max_{(\theta_1, \theta_2)} \log \mathcal{L}(\theta_1, \theta_2 | D_1, D_2, U) \quad (7)$$

This can be done with an EM algorithm, whose steps are summarized in Algorithm 1. First, the MaxEnt parameters, θ_1 and θ_2, are estimated from just the labeled data. Then, in the E-step, the classifiers, based on current values of θ_1 and θ_2, compute $p(y_i|x_i)$ for each labeled example and assign probabilistically-weighted class labels to each unlabeled example. Next, in the M-step, the parameters, θ_1 and θ_2, are updated using both the original labeled data (D_1 and D_2) and the newly labeled data U. These last two steps are iterated until convergence or a predefined iteration limit T.

Algorithm 1. The MaxEnt-based EM Algorithm for Multilingual Sentiment Classification

Input: Labeled data D_1 and D_2
Unlabeled parallel data U

Output: Two monolingual MaxEnt classifiers with parameters θ_1^* and θ_2^*, respectively

1. **Train two initial monolingual models**
   train and initialize $\theta_1^{(0)}$ and $\theta_2^{(0)}$ on the labeled data
2. **Jointly optimize two monolingual models**
   for $t = 1$ **to** T **do** // T: number of iterations
   E-Step:
   Compute $p(y|x)$ for each example in D_1, D_2 and U based on $\theta_1^{(t-1)}$ and $\theta_2^{(t-1)}$;
   Compute the expectation of the log likelihood with respect to $p(y|x)$;
   M-Step:
   Find $\theta_1^{(t)}$ and $\theta_2^{(t)}$ by maximizing the regularized joint log likelihood;
   Convergence:
   If the increase of the joint log likelihood is sufficiently small, break;
   end for
3. Output θ_1^* as $\theta_1^{(t)}$s, and θ_2^* as $\theta_2^{(t)}$

In the M-step, we can optimize the regularized joint log likelihood using any gradient-based optimization techniques (Malouf, 2002). The gradient for Formula 3 based on Formula 4 is shown in Appendix A, and those for Formula 5 and 6 can be derived similarly. In our experiments, we use the L-BFGS algorithm (Liu et al., 1989) and run EM until the change in regularized joint log likelihood is less than 1e-5 or we reach 100 iterations.[3]

[2] The probability should be rescaled within the range of [0, 1], where 0.5 means that we are completely unsure if the sentences are translations of each other or not, and only those translation pairs with a probability larger than 0.5 are meaningful for our purpose.

[3] Since the EM-based algorithm may find a local maximum of the objective function, the initialization of the parameters is important. The empirical experiments show that it usually finds an effective maximum by initializing the parameters as those learned from the labeled data, but the performance would be quite worse if we initialize all the parameters with 0 or 1.

3.4 Pseudo-Parallel Labeled and Unlabeled Data

We also consider the case where a parallel corpus is not available: to obtain a pseudo-parallel corpus U (i.e. sentences in one language with their corresponding automatic translations), we use an automatic machine translation system (e.g. Google machine translation [4]) to translate unlabeled in-domain data from L_1 to L_2 or vice versa.

Since previous work (Banea et al., 2008; 2010; Wan, 2009) has shown that it could be useful to automatically translate the labeled data from the source language into the target language, we can further incorporate such translated labeled data into the joint model by adding the following component into Formula 6:

$$\lambda_3 \sum_{\bar{v}=1}^{2} \sum_{i=1}^{l_{\bar{v}}} \log p\left(y_i^{\bar{v}} \middle| x_i^{\bar{v}^*}; \theta_v\right) \quad (8)$$

where $\bar{v}$ is the alternative class of v, $x_i^{\bar{v}^*}$ is the automatically translated example from $x_i^{\bar{v}}$; and $\lambda_3 \geq 0$ is a constant that controls the weight of the translated labeled data.

4 Experimental Setup

4.1 Data Sets and the Preprocessing

The following labeled datasets are used in our experiments:

MPQA (Labeled English Data): The Multi-Perspective Question Answering (MPQA) corpus (Wiebe et al., 2005) consists of newswire documents manually annotated with phrase-level subjectivity information. We extract all sentences containing strong (i.e. intensity is *medium* or higher), sentiment-bearing (i.e. polarity is *positive* or *negative*) expressions following Choi and Cardie (2008). Sentences with both positive and negative strong expressions are then discarded, and the polarity of each remaining sentence is set to that of its sentiment-bearing expression(s).

NTCIR-EN (Labeled English Data) and NTCIR-CH (Labeled Chinese Data): The NTCIR Opinion Analysis task (Seki et al., 2007; 2008) provides sentiment-labeled news data in Chinese, Japanese and English. Only those sentences with a polarity label (positive or negative) agreed by at least two annotators are extracted. We use the Chinese data from NTCIR-6 as our Chinese labeled data. Since far fewer sentences in the English data passed the annotator agreement filter, we combined the English data from NTCIR-6 and NTCIR-7. The Chinese sentences were segmented using the Stanford Chinese word segmenter (Tseng et al., 2005).

The number of sentences in each of these datasets is shown in Table 1. In our experiments, we evaluate two settings of the data: 1) MPQA+NTCIR-CH, and 2) NTCIR-EN+NTCIR-CH. In each setting, the English labeled data constitutes D_1 and the Chinese labeled data, D_2.

	MPQA	NTCIR-EN	NTCIR-CH
Positive	1,471(30%)	528 (30%)	2,378(55%)
Negative	3,487(70%)	1,209 (70%)	1,916 (45%)
Total	4,958	1,737	4,294

Table 1: Sentence Numbers of the Labeled Data

Unlabeled Parallel Text and its Preprocessing

For the unlabeled parallel text, we use the ISI Chinese-English parallel corpus (Munteanu and Marcu, 2005), which was extracted automatically from news articles published by Xinhua News Agency in Chinese Gigaword (2nd Edition) and English Gigaword (2nd Edition). Because the sentence pairs in the ISI corpus are quite noisy, we rely on Giza++ (Och and Ney, 2003) to get the new translation probability for each sentence pair, and we choose only the top 100,000 pairs with the highest translation probabilities.[5]

We also try to remove neutral sentences from the parallel data since they can introduce noise into our model, which deals only with positive and negative examples. We train one MaxEnt classifier on both Chinese and English labeled data for each data setting above, and then classified each unlabeled sentence pair by combining the two sentences in each pair into one. Then we choose the most confidently predicted 10,000 positive and 10,000 negative pairs to constitute the unlabeled parallel corpus U for each data setting.

[4] http://translate.google.com/

[5] We removed the sentence pairs with an original confidence score (given in the corpus) smaller than 0.98, and also removed the pairs that are too long (with more than 60 characters in one sentence) to facilitate Giza++. We first get the translation probabilities for both directions (i.e. Chinese to English and English to Chinese) with Giza++, take log of the product of two probabilities, and then divide it by the sum of lengths of the two sentences in each pair.

4.2 Baseline Methods

In our experiments, the proposed joint model is compared with the following baseline methods:

MaxEnt: This method applies MaxEnt classifier for each language on the monolingual labeled data, and the unlabeled data is not used.

SVM: This method applies an SVM classifier for each language with the monolingual labeled data, and the unlabeled data is not used. SVM-light (Joachims, 1999a) is used for all the SVM-related experiments.

Monolingual TSVM (TSVM-M): This method learns two transductive SVM (TSVM) classifiers with the monolingual labeled training data and the monolingual unlabeled data for each language.

Bilingual TSVM (TSVM-B): This method learns one TSVM classifier based on the labeled training data in two languages, as well as the unlabeled sentences by combining the two sentences in each unlabeled pair into one. The method is supposed to perform better than TSVM-M since the combined unlabeled sentences could be more helpful than the monolingual sentences.

Co-Training on SVM (Co-SVM): This method applies the SVM-based co-training with both the labeled training data and the unlabeled parallel data following Wan (2009). First, two monolingual SVM classifiers are built based on only the corresponding labeled data, and then they are bootstrapped by adding the most confident predicted examples from the unlabeled data into the training set. We run bootstrapping for 100 iterations. In each iteration, we select the most confidently predicted 50 positive and 50 negative sentences from each classifier, and take the union of the 200 sentence pairs as the newly labeled training data (Note the examples with conflicting labels are not included).

5 Results and Analysis

In our experiments, the methods are tested in the two data settings with the corresponding unlabeled parallel corpus as mentioned above.[6] We do 5-fold cross-validation to get the average accuracy (also MicroF1 in this case) and MacroF1 as the overall metrics for each method. Unigrams are used as binary features for all models, as Pang et al. (2002) showed that binary features perform better than frequency features in sentiment classification. The weights for unlabeled data and regularization, λ_1 and λ_2, are set to 1, unless otherwise stated. Later, we will show the proposed approach performs well with a wide range of parameter values.[7]

5.1 Method Comparison

We first compare the proposed joint model (**Joint**) with the baselines. Table 2 shows the comparison. Seen from the table, the proposed approach outperforms all the five baseline methods in terms of both accuracy and MacroF1 on both English and Chinese in both of the two data settings.[8] By making use of the unlabeled parallel data, our proposed approach improves the accuracy, compared to MaxEnt, by 8.12% (or 33.27% error reduction) on English and by 3.44% (or 16.92% error reduction) on Chinese in the first setting, and by 5.07% (or 19.67% error reduction) on English and by 3.87% (or 19.4% error reduction) on Chinese in the second setting.

Among the baselines, the best one is Co-SVM; TSVMs do not always improve performance on our test dataset by using the unlabeled data, compared to SVM alone; and TSVM-B outperforms TSVM-M in most cases except on Chinese in the second setting. MPQA is more difficult in general compared to the NTCIR data. Without unlabeled parallel data, the performances on the Chinese data are better than those on the English data, which is consistent with the results reported in NTCIR-6 (Seki et al., 2007).

Overall, the unlabeled parallel data can be beneficial to the classification accuracy for both languages by using our proposed joint model and Co-SVM. The joint model is more suitable for making use of the unlabeled parallel data than Co-SVM or TSVMs, and the reason is that the proposed approach tries to jointly optimize the two monolingual models with soft (probabilistic)

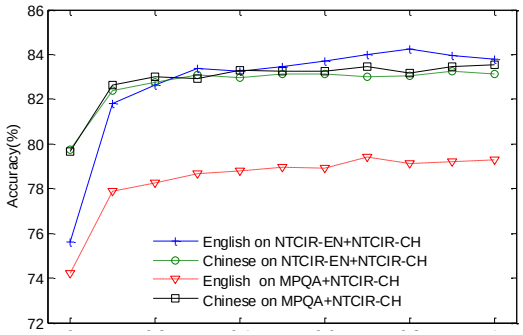
[6] The results reported in this section are all done with Formula 4. The preliminary experiments show that Formula 5 does not significantly improve the performance in our case, which is reasonable since we choose only the sentence pairs with the highest translation probabilities to be our unlabeled data as mentioned in Section 4.1.

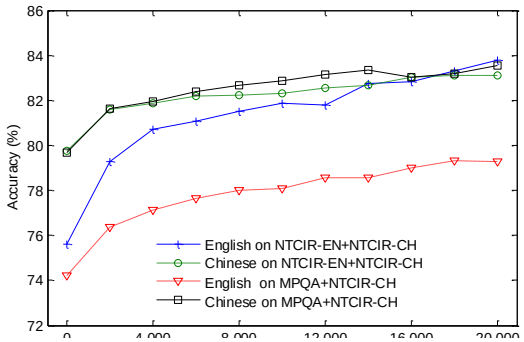
[7] The code is at http://sites.google.com/site/lubin2010.

[8] The significant tests are done by using paired t-tests with $p<0.05$: € denotes statistically significant compared to the corresponding performance of MaxEnt; * denotes that statistically significant compared to SVM; and ┌ denotes statistically significant compared to Co-SVM.

Setting 1: NTCIR-EN+NTCIR-CH				**Setting 2**: MPQA+NTCIR-CH			
Accuracy		MacroF1		Accuracy		MacroF1	
English	Chinese	English	Chinese	English	Chinese	English	Chi
75.59	79.67	66.61*	79.34	74.22	79.67	65.09*	
76.34	81.02	61.12	80.75$^{€}$	76.74$^{€}$	81.02	61.35	80
73.46	80.21	55.33	79.99	72.89	81.14	52.82	79
78.36	81.60$^{€}$	65.53	81.42	76.42$^{€}$	78.51	61.66	78
82.44$^{€*}$	82.79$^{€}$	72.61$^{€*}$	82.67$^{€*}$	78.18$^{€*}$	82.63$^{€*}$	68.03$^{€*}$	82.
83.71$^{€*}$	**83.11**$^{€*}$	**75.89**$^{€*Γ}$	**82.97**$^{€*}$	**79.29**$^{€*Γ}$	**83.54**$^{€*}$	**72.58**$^{€*Γ}$	**83.**

Table 2: Comparison of Results





assignments of the unlabeled instances to classes in each iteration, instead of hard assignments in Co-SVM or TSVMs. Although English sentiment classification alone is more difficult than Chinese on the datasets, we can obtain greater performance gains on English by exploiting unlabeled parallel data and the Chinese labeled data.

5.2 Varying the Weight and Amount of Unlabeled Data

Fig. 1 shows the accuracy curve of the proposed approach for the two data settings by varying the weight for the unlabeled data, λ_1, from 0 to 1. When λ_1 is set to 0, the joint model is degenerated to two MaxEnt models with only the labeled data.

We can see that the performance gain of the proposed approach is quite remarkable even when λ_1 is set to 0.1, and the performance is stable or changes slightly after it reaches 0.4. Although MPQA is more difficult, in general, compared to the NTCIR data, we can still steadily improve its performance with unlabeled parallel data. Overall, the proposed approach performs quite well with a wide range of parameter values of λ_1.

Fig. 2 shows the accuracy curve of the proposed approach for the two data settings by varying the amount of unlabeled data from 0 to 20,000. We can see that the performance of the proposed approach improves steadily by adding more and more unlabeled data. However, even with only 2,000 unlabeled sentence pairs, the proposed approach still produces remarkable performance gains. MPQA still is more difficult compared to the NTCIR data.

5.3 Results on Pseudo-Parallel Unlabeled Data

As discussed in Section 3.4, we generate pseudo-parallel data by translating the monolingual sentences in each setting using Google machine translation. Fig. 3 and 4 show the performances of our model on the pseudo-parallel data and the real parallel data, in the two settings, respectively. The EN->CH pseudo-parallel data consists of the English unlabeled data and its automatic Chinese translation, and vice versa.

Although the improvements are not as significant as those with parallel data, we can still obtain improvement with the pseudo-parallel data, especially in the first setting. The difference between using parallel versus pseudo-parallel data is around 2-4% in Fig 3 and 4, which is reasonable since the quality of pseudo-parallel data is not as good as that of the parallel data. Therefore, the performance with pseudo-parallel data is better with a small weight λ_1(e.g. 0.1) at some cases.

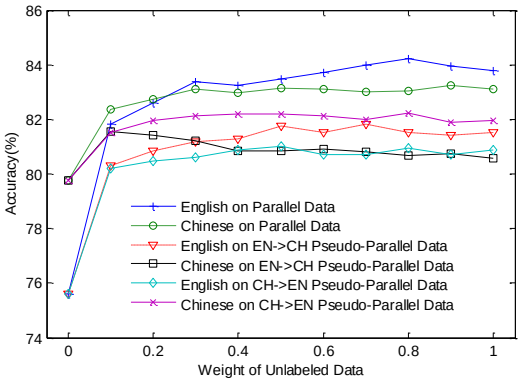


Fig. 3. Accuracy with Pseudo-Parallel Unlabeled Data in **Setting 1**

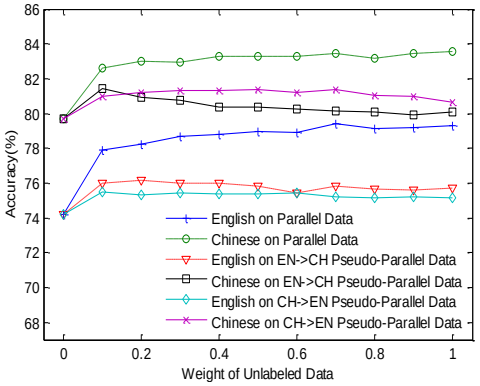


Fig. 4. Accuracy with Pseudo-Parallel Unlabeled Data in **Setting 2**

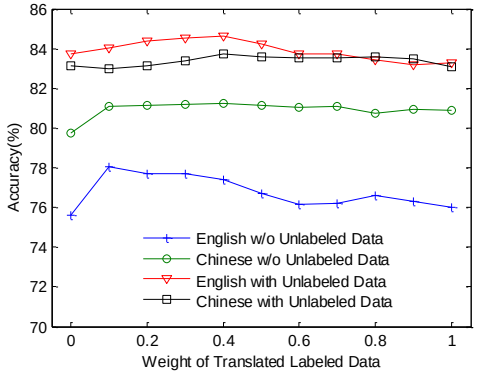


Fig. 5. Accuracy with Pseudo-Parallel Labeled Data in **Setting 1**

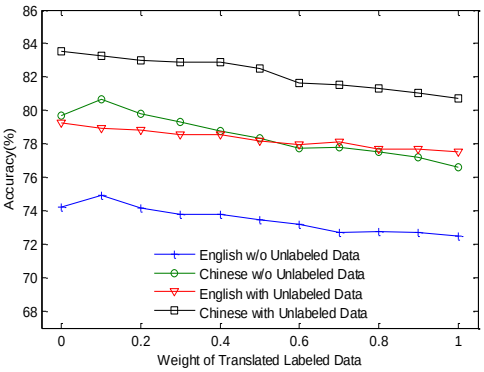


Fig. 6. Accuracy with Pseudo-Parallel Labeled Data in **Setting 2**

5.4 Adding Pseudo-Parallel Labeled Data

In this section, we investigate how adding automatically translated labeled data might influence the performance as mentioned in Section 3.4. We first use only the translated labeled data to train classifiers, and then immediately classify the test data. The average accuracies in setting 1 are 66.61% and 63.11% on English and Chinese, respectively; while the accuracies in setting 2 are 58.43% and 54.07% on English and Chinese, respectively. This result is reasonable because of the language gap between the original language and the translated language. Another reason is that the class distributions between the English labeled data and the Chinese are quite different (30% vs. 55% for positive as shown in Table 1).

Fig. 5 and 6 show the accuracies by varying the weight of the translated labeled data when adding them into the labeled data with and without the unlabeled parallel data. From Fig. 5 for setting 1, we can see that the translated data could be helpful given the labeled data and even the unlabeled data, as long as λ_3 is small; while in Fig. 6, the translated data decreases the performance in most cases for setting 2. The possible reason is that in the first data setting, the NTCIR English data covers the same topics as the NTCIR Chinese data and thus direct translation is helpful, while the English and Chinese topics are quite different in the second data setting, and thus direct translation hurts the performance given the existing labeled data in each language.

5.5 Discussion

To further understand what contributions our proposed approach makes to the performance gain, we look inside the parameters in the MaxEnt models learned before and after adding the parallel unlabeled data. Table 3 shows the features which are already in the model learned from the labeled data, but have the largest weight change after adding the parallel data; and Table 4 shows the newly learned features from the unlabeled data with the largest weights.

	Word	Weight		
		Before	After	Change
Positive	important	0.452	1.659	1.207
	cooperation	0.325	1.492	1.167
	support	0.533	1.483	0.950
	importance	0.450	1.193	0.742
	agreed	0.347	1.061	0.714
Negative	difficulties	0.018	0.663	0.645
	not	0.202	0.844	0.641
	never	0.245	0.879	0.634
	germany	0.035	0.664	0.629
	taiwan	0.590	1.216	0.626

Table 3. Original Features with Largest Weight Change

Positive		Negative	
Word	Weight	Word	Weight
friendly	0.701	german	0.783
principles	0.684	arduous	0.531
hopes	0.630	oppose	0.511
hoped	0.553	administrations	0.431
cooperative	0.552	oau[9]	0.408

Table 4. New Features Learned from Unlabeled Data

From Table 3 and 4, [10] we can see the weight changes of the original features and the new features are quite reasonable, e.g. the top words in the positive class are obviously positive and the proposed approach gives them higher weights. For some specific words (i.e. germany, taiwan, german, etc.), the labeled and unlabeled data have some negative news about them.

We also examine the process of joint training by checking the performance on test data and the agreement of the two monolingual models on the unlabeled parallel data in both settings. The average agreements across 5 folds are 85.06% and 73.87% in setting 1 and 2 respectively before the joint training, and increase to 100% and 99.89% respectively after 100 iterations of joint training. Although the average agreements have already increased to 99.50% and 99.02% in setting 1 and 2 respectively after 30 iterations, the performances on the test sets are still steadily improving in both settings until around 50-60 iterations, and then become relatively stable after that.

Those sentence pairs in setting 2 which are still disagreed by the two monolingual models after 100 iterations of joint training are not quite parallel, e.g. the sentence pair below:

English: The two sides attach great importance to international cooperation on protection and promotion of human rights .

Chinese: 双方认为,在人权问题上不能采取“双重标准”,反对在国际关系中利用人权问题施压。(Both sides agree that double standards on the issue of human rights are to be avoided, and oppose to using pressure on human rights issues in international relations.)

Since the two sentences are talking about *human rights* from very different perspectives, it is reasonable that even after joint training, the two monolingual models still classify them into different polarities (i.e. positive for the English sentence and negative for the Chinese sentence).

6 Conclusion

In this paper, we study bilingual sentiment classification and propose a joint model to simultaneously learn better monolingual sentiment classifiers for each language by exploiting an unlabeled parallel corpus together with the labeled data in each language. Our experiments show that the proposed approach can significantly improve sentiment classification for both languages. Moreover, the proposed approach continues to produce (albeit smaller) performance gains when employing pseudo-parallel data from machine translation engines.

In future work, we would like to apply the joint learning idea to other frameworks (e.g. SVM), and to extend the proposed model to handle word-level parallel information, e.g. bilingual dictionaries or word alignment information. Another issue is to investigate how to improve multilingual sentiment analysis by exploiting comparable corpora.

Acknowledgments

We thank Shuo Chen, Long Jiang, Thorsten Joachims, Lillian Lee, Myle Ott, Yan Song, Xiaojun Wan, Ainur Yessenalina, Jingbo Zhu and the anonymous reviewers for many useful comments and discussion. This work was supported in part by National Science Foundation Grants BCS-0904822, BCS-0624277, IIS-0968450; and by a gift from Google. Chenhao Tan is supported by NSF (DMS-0808864), ONR (YIP-N000140910911), and a grant from Microsoft.

[9] It is an abbreviation for the Organization of African Unity.

[10] The features and weights in Table 3 and 4 are extracted from the English model in the first fold of setting 1.

References

Massih-Reza Amini, Cyril Goutte, and Nicolas Usunier. 2010. Combining coregularization and consensus-based self-training for multilingual text categorization. In *Proceeding of SIGIR'10*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of COLING'10*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP'08*.

Adam L. Berger, Stephen A. Della Pietra and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1).

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspond-dence learning. In *Proceedings of EMNLP'06*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT'98*.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised Latent Dirichlet Allocation. In *Proceedings of EMNLP'10*.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of IJCAI'07*.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of CoNLL'10*.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of EMNLP'08*.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of EMNLP'08*.

Wei Gao, John Blitzer, Ming Zhou, and Kam-Fai Wong. 2009. Exploiting bilingual information to improve web search. In *Proceedings of ACL/IJCNLP'09*.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of AAAI'04*.

Ido Dagan, and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus, Computational Linguistics, 20(4): 563-596.

Thorsten Joachims. 1999a. Making Large-Scale SVM Learning Practical. In: *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola (ed.), MIT Press.

Thorsten Joachims. 1999b. Transductive inference for text classification using support vector machines. In *Proceedings of ICML'99*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of NAACL'06*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, (45): 503–528.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of CoNLL'02*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL'07*.

Dragos S. Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4): 477–504.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *Proceedings of NAACL/HLT '10*.

Kamal Nigam, Andrew K. Mccallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2): 103–134.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-51.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, Now Publishers.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP'02*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of ACL'10*.

Adwait Ratnaparkhi. 1997. A simple introduction to maximum entropy models for natural language processing. Technical Report 97-08, University of Pennsylvania.

Julia M. Schulz, Christa Womser-Hacker, and Thomas Mandl. 2010. Multilingual corpus development for opinion mining. In *Proceedings of LREC'10*.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-His Chen, and Noriko Kando. 2008. Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of the NTCIR-7 Workshop*.

Yohei Seki, David K. Evans, Lun-Wei Ku, Le Sun, Hsin-His Chen, Noriko Kando, and Chin-Yew Lin. 2007. Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of the NTCIR-6 Workshop*.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. 2005. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML'05*.

Noah A. Smith. 2006. Novel estimation methods for unsupervised discovery of latent structure in natural language text. Ph.D. thesis, Department of Computer Science, Johns Hopkins University.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A conditional random field word segmenter. In *Proeedings of the 4th SIGHAN Workshop*.

Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In *Proceedings of ACL'02*.

Xiaojun Wan. 2008. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In *Proceedings of EMNLP'08.*

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL/AFNLP'09*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2- 3): 165-210.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction, In *Proceedings of ACL'10*.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of ACL/IJCNLP'09*.

Xiaojin Zhu and Andrew B. Goldberg. 2009. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.

Appendix A. Formula Deduction

In this appendix, we derive the gradient for the objective function in Formula 3, which is used in parameter estimation. As mentioned in Section 3.3, the parameters can be learned by finding:

$$\begin{aligned}(\theta_1^*, \theta_2^*) &= \underset{(\theta_1,\theta_2)}{\operatorname{argmax}} \mathcal{L}(\theta_1, \theta_2 | D_1, D_2, U) \\ &= \underset{(\theta_1,\theta_2)}{\operatorname{argmax}} \log \mathcal{L}(\theta_1, \theta_2 | D_1, D_2, U) \\ &= \underset{(\theta_1,\theta_2)}{\operatorname{argmax}} \{\log p(Y_1|X_1;\theta_1) p(Y_2|X_2;\theta_2) \\ &\quad + \textstyle\sum_{i=1}^{u} \log p(y_i^{1'}, y_i^{2'} | x_i^{1'}, x_i^{2'}; \theta_1, \theta_2)\}\end{aligned}$$

Since the first term on the right hand side is just the expression for the standard MaxEnt problem, we will focus on the gradient for the second term, and denote $\log p(y_i^{1'}, y_i^{2'} | x_i^{1'}, x_i^{2'}; \theta_1, \theta_2)$ as $(*)$.

Let $v \in \{1,2\}$ denote L_1 or L_2, and θ_k^v be the k-th weight in the vector θ_v. For brevity, we drop the $'$ in the above notations, and write x_i^v to denote $x_i^{v'}$. Then the partial derivative of (*) based on Formula 4 with respect to θ_k^v is as follows:

$$\frac{\partial(*)}{\partial \theta_k^v} = \frac{\sum_{y_i^*} p\left(y_i^* | x_i^{\bar{v}}; \theta_{\bar{v}}\right) \frac{\partial}{\partial \theta_k^v} p(y_i^* | x_i^v; \theta_v)}{\sum_{y_i'} p(y_i' | x_i^v; \theta_v) p(y_i' | x_i^{\bar{v}}; \theta_{\bar{v}})} \quad (1)$$

Further, we can get:

$$\begin{aligned}&\frac{\partial}{\partial \theta_k^v} p(y_i^* | x_i^v; \theta_v) = \frac{\partial}{\partial \theta_k^v} \frac{\exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i^*)\right)}{\sum_{y_i'} \exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i')\right)} \\ &= \frac{\exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i^*)\right)}{\sum_{y_i'} \exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i')\right)} f_k^v(x_i^v, y_i^*) - \\ &\quad \frac{\exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i^*)\right)}{\left[\sum_{y_i'} \exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i')\right)\right]^2} \sum_{y_i'} \left\{\exp\left(\vec{\theta}_v \vec{f}(x_i^v, y_i')\right) f_k^v(x_i^v, y_i')\right\} \\ &= p(y_i^* | x_i^v; \theta_v) \left\{f_k^v(x_i^v, y_i^*) - \textstyle\sum_{y_i'} p(y_i' | x_i^v; \theta_v) f_k^v(x_i^v, y_i')\right\} \quad (2)\end{aligned}$$

Merge (2) into (1), we get:

$$\begin{aligned}\frac{\partial(*)}{\partial \theta_k^v} &= \frac{1}{\sum_{y_i'} p(y_i' | x_i^v; \theta_v) p(y_i' | x_i^{\bar{v}}; \theta_{\bar{v}})} \sum_{y_i^*} \{p\left(y_i^* | x_i^{\bar{v}}; \theta_{\bar{v}}\right) p(y_i^* | x_i^v; \theta_v) \\ &\quad [f_k^v(x_i^v, y_i^*) - \textstyle\sum_{y_i'} p(y_i' | x_i^v; \theta_v) f_k^v(x_i^v, y_i')]\} \\ &= \textstyle\sum_{y_i^*} p\left(y_i^* | x_i^1; \theta_1\right) p\left(y_i^* | x_i^2; \theta_2\right) f_k^v(x_i^v, y_i^*) - \\ &\quad \textstyle\sum_{y_i'} p(y_i' | x_i^v; \theta_v) f_k^v(x_i^v, y_i') \\ &= \textstyle\sum_{y_i^*} f_k^v(x_i^v, y_i^*) \{p(y_i^* | x_i^1; \theta_1) p(y_i^* | x_i^2; \theta_2) - p(y_i^* | x_i^v; \theta_v)\}\end{aligned}$$