Latent Sentiment Model for Weakly-Supervised Cross-Lingual Sentiment Classification

Yulan He

Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes MK7 6AA, UK
{y.he}@open.ac.uk

Abstract. In this paper, we present a novel weakly-supervised method for cross-lingual sentiment analysis. In specific, we propose a latent sentiment model (LSM) based on latent Dirichlet allocation where sentiment labels are considered as topics. Prior information extracted from English sentiment lexicons through machine translation are incorporated into LSM model learning, where preferences on expectations of sentiment labels of those lexicon words are expressed using generalized expectation criteria. An efficient parameter estimation procedure using variational Bayes is presented. Experimental results on the Chinese product reviews show that the weakly-supervised LSM model performs comparably to supervised classifiers such as Support vector Machines with an average of 81% accuracy achieved over a total of 5484 review documents. Moreover, starting with a generic sentiment lexicon, the LSM model is able to extract highly domain-specific polarity words from text.

Keywords: Latent sentiment model (LSM), cross-lingual sentiment analysis, Generalized expectation, latent Dirichlet allocation.

1 Introduction

The objective of sentiment analysis is to automatically extract opinions, emotions and sentiments in text. It allows us to track attitudes and feelings on the web. Research in sentiment analysis has mainly focused on the English language. Little work has been carried out in other languages partly due to the lack of resources, such as subjectivity lexicons consisting of a list of words marked with their respective polarity (positive, negative or neutral) and manually labeled subjectivity corpora with documents labeled with their polarity.

Pilot studies on cross-lingual sentiment analysis utilize machine translation to perform sentiment analysis on the English translation of foreign language text [10, 1, 2, 17]. These supervised learning algorithms suffer from the generalization problem since annotated data (either annotated English corpora or the translated corpora generated by machine translation) are required for classifier training. As such, they often fails when there is a domain mismatch between the source and target languages. Recent effort has been made to exploit bootstrapping-style approaches for weakly-supervised sentiment classification in languages other than English [19, 18, 12]. Other approaches use ensemble techniques by either combining lexicon-based and corpus-based algorithms [15] or combining sentiment classification outputs from different experimental settings [16]. Nevertheless, all these approaches are either complex or require careful tuning of domain and data specific parameters.

This paper proposes a weakly-supervised approach for cross-lingual sentiment classification by incorporating lexical knowledge obtained from available English sentiment lexicons through machine translation. Preferences on expectations of sentiment labels of those lexicon words are expressed using generalized expectation criteria [9] and are used to modify the latent Dirichlet allocation (LDA) model objective function for model learning, which we named as latent sentiment model (LSM). The proposed approach performs sentiment analysis without the use of labeled documents. In addition, it is simple and computationally efficient; rendering more suitable for online and real-time sentiment classification from the Web.

Incorporating sentiment prior knowledge into LDA model for sentiment analysis has been previously studied in [8, 7] where the LDA model has been modified to jointly model sentiment and topic. However their approach uses the sentiment prior information in the Gibbs sampling inference step that a sentiment label will only be sampled if the current word token has no prior sentiment as defined in a sentiment lexicon. This in fact implies a different generative process where many of the l's are observed. The model is no longer "latent". Our proposed approach incorporate sentiment prior knowledge in a more principled way that we express preferences on expectations of sentiment labels of the lexicon words from a sentiment lexicon using generalized expectation criteria and essentially create an informed prior distribution for the sentiment labels. This would allow the model to actually be latent and would be consistent with the generative story.

We have explored several commonly used English sentiment lexicons and conducted extensive experiments on the Chinese reviews of four different product types. The empirical results show that the LSM model, despite using no labeled documents, performs comparably to the supervised classifiers such as Support Vector Machines (SVMs) trained from labeled corpora. Although this paper primarily studies sentiment analysis in Chinese, the proposed approach is applicable to any other language so long as a machine translation engine is available between the selected language and English.

The remainder of the paper is structured as follows. Related work on cross-lingual and weakly-supervised sentiment classification in languages other than English are discussed in Section 2. Existing algorithms on incorporating supervised information into LDA model learning are also reviewed in this section. The proposed LSM model and its inference and training procedures are presented in Section 3. The experimental setup and results of sentiment classification on the Chinese reviews of four different products are presented in Section 4 and 5 respectively. Finally, Section 6 concludes the paper.

2 Related Work

Early work on cross-lingual sentiment analysis rely on English corpora for subjectivity classification in other languages. For example, Mihalcea et al. [10] make use of a bilingual lexicon and a manually translated parallel text to generate the resources to build subjectivity classifiers based on SVMs and Naïve Bayes (NB) in a new language;

Banea et al. [1] use machine translation to produce a corpus in a new language and train SVMs and NB for subjectivity classification in the new language. Bautin et al. [2] also utilize machine translation to perform sentiment analysis on the English translation of a foreign language text. More recently, Wan [17] proposed a co-training approach to tackle the problem of cross-lingual sentiment classification by leveraging an available English corpus for Chinese sentiment classification. The major problem of these cross-lingual sentiment analysis algorithms is that they all utilize supervised learning to train sentiment classifiers from annotated English corpora (or the translated target language corpora generated by machine translation). As such, they cannot be generalized well when there is a domain mismatch between the source and target language.

Recent efforts have also been made for weakly-supervised sentiment classification in languages other than English. Zagibalov and Carroll [19, 18] starts with a one-word sentiment seed vocabulary and use iterative retraining to gradually enlarge the seed vocabulary by adding more sentiment-bearing lexical items based on their relative frequency in both the positive and negative parts of the current training data. Sentiment direction of a document is then determined by the sum of sentiment scores of all the sentiment-bearing lexical items found in the document. Qiu et al. [12] also uses a lexicon-based iterative process as the first phase to iteratively enlarge an initial sentiment dictionary. Documents classified by the first phase are taken as the training set to train the SVMs which are subsequently used to revise the results produced by the first phase. Wan [16] combined sentiment scores calculated from Chinese product reviews using the Chinese HowNet sentiment dictionary[1] and from the English translation of Chinese reviews using the English MPQA subjectivity lexicon[2]. Various weighting strategies were explored to combine sentiment classification outputs from different experimental settings in order to improve classification accuracy. Nevertheless, all these weakly-supervised sentiment classification approaches are rather complex and require either iterative training or careful tuning of domain and data specific parameters, and hence unsuitable for online and real-time sentiment analysis in practical applications.

In recent years, there have been increasing interests in incorporating supervised information into LDA model learning. Blei and McAuliffe [3] proposed supervised LDA (sLDA) which uses the empirical topic frequencies as a covariant for a regression on document labels such as movie ratings. Mimno and McCallum [11] proposed a Dirichlet-multinomial regression which uses a log-linear prior on document-topic distributions that is a function of observed features of the document, such as author, publication venue, references, and dates. DiscLDA [6] and Labeled LDA [13] assume the availability of document class labels and utilize a transformation matrix to modify Dirichlet priors. DiscLDA introduces a class-dependent linear transformation to project a K-dimensional (K latent topics) document-topic distribution into a L-dimensional space (L document labels), while Labeled LDA simply defines a one-to-one correspondence between LDA's latent topics and document labels. Our work differs from theirs in that we use word prior sentiment as supervised information and modify the LDA objective function by adding the generalized expectation criteria terms.

[1] http://www.keenage.com/download/sentiment.rar
[2] http://www.cs.pitt.edu/mpqa/

3 Latent Sentiment Model

Unlike existing approaches, we view sentiment classification as a generative problem that when an author writes a review document, he/she first decides on the overall sentiment or polarity (positive, negative, or neutral) of a document, then for each sentiment, decides on the words to be used. The LSM model, as shown in Figure 1, can be treated as a special case of LDA where a mixture of only three sentiment labels are modeled, i.e. positive, negative and neutral.

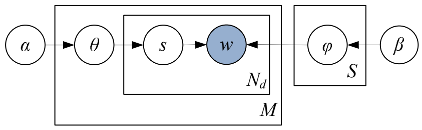


Fig. 1. Latent sentiment model.

Assuming that we have a total number of S sentiment labels $\mathcal{S} = \{neutral, positive, negative\}$; a corpus with a collection of M documents is denoted by $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$, where the bold-font variables denote the vectors; each document in the corpus is a sequence of N_d words denoted by $\mathbf{w} = (w_1, w_2, ..., w_{N_d})$, and each word in the document is an item from a vocabulary index with V distinct terms denoted by $\{1, 2, ..., V\}$. The generative process is to first draw $\boldsymbol{\varphi}_s \sim \text{Dir}(\boldsymbol{\beta})$, then for each document $d \in [1, M]$, choose a distribution $\theta_d \sim \text{Dir}(\alpha)$, and for each of the N_d word position w_t, sample a sentiment label $s_t \sim \text{Multinomial}(\theta_d)$ and choose a word $w_t \sim \text{Multinomial}(\boldsymbol{\varphi}_{s_t})$.

Letting $\Lambda = \{\alpha, \beta\}$, we obtain the marginal distribution of a document $\mathbf{w}$ by integrating over θ and φ and summing over s:

$$P(\mathbf{w}|\Lambda) = \int\int P(\theta;\alpha)\prod_{s=1}^{S} P(\varphi_s;\beta)\prod_{t=1}^{N_d}\sum_{s_t} p(s_t|\theta)P(w_t|s_t,\varphi_{s_t})d\theta d\varphi \quad (1)$$

Taking the product of marginal probabilities of documents in a corpus gives us the probability of the corpus.

$$P(\mathcal{D}|\Lambda) = \prod_{d=1}^{M} P(\mathbf{w}_d|\Lambda) \quad (2)$$

Assume we have some labeled features where words are given with their prior sentiment orientation, we could construct a set of real-valued features of the observation to expresses some characteristic of the empirical distribution of the training data that should also hold of the model distribution.

$$f_{jk}(w,s) = \sum_{d=1}^{M}\sum_{t=1}^{N_d} \delta(s_{d,t} = j)\delta(w_{d,t} = k) \quad (3)$$

where $\delta(x)$ is an indicator function which takes a value of 1 if x is true, 0 otherwise. Equation 3 calculates how often feature k and sentiment label j co-occur in the corpus.

We define the expectation of the features as

$$E_{\Lambda}[\mathbf{f}(w,s)] = E_{\tilde{P}(w)}[E_{P(w|s;\Lambda)}[\mathbf{f}(w,s)]] \tag{4}$$

where $\tilde{P}(w)$ is the empirical distribution of w in document corpus $\mathcal{D}$, and $P(w|s;\Lambda)$ is a conditional model distribution parameterized at Λ.

$E_{\Lambda}[\mathbf{f}(w,s)]$ is a matrix of size $S \times K$ where S is the total number of sentiment labels and K is the total number of features or constraints used in model learning. The jkth entry denotes the expected number of times that feature k is assigned with label j.

We define a criterion that minimizes the KL divergence of the expected label distribution and a target expectation $\hat{\mathbf{f}}$, which is essentially an instance of generalized expectation criteria that penalizes the divergence of a specific model expectation from a target value.

$$G(E_{\Lambda}[\mathbf{f}(w,s)]) = KL(\hat{\mathbf{f}}||E_{\Lambda}[\mathbf{f}(w,s)]) \tag{5}$$

We can use the target expectation $\hat{\mathbf{f}}$ to encode human or task prior knowledge. For example, the word "*excellent*" typically represent a positive orientation. We would expect that this word more likely appears in positive documents. In our implementation, we adopted a simple heuristic approach [14, 4] that a majority of the probability mass for a feature is distributed uniformly among its associated labels, and the remaining probability mass is distributed uniformly among the other non-associated label(s). As we only have three sentiment labels here, the target expectation of a feature having its prior polarity (or associated sentiment label) is 0.9 and 0.05 for its non-associated sentiment labels.

The above encodes word sentiment prior knowledge in the form of $\hat{P}(s|w)$. However, the actual target expectation used in our approach is $\hat{P}(w|s)$. We could perform the following simple transformation:

$$\hat{P}(w|s) = \frac{\hat{P}(s|w)P(w)}{P(s)} \propto \hat{P}(s|w)\tilde{P}(w) \tag{6}$$

by assuming that the prior probability of w can be obtained from the empirical distribution of w in document corpus $\mathcal{D}$, and the prior probability of the three sentiment labels are uniformly distributed in the corpus.

We augment the likelihood maximization by adding the generalized expectation criteria objective function terms.

$$\mathcal{O}(\mathcal{D}|\Lambda) = \log P(\mathcal{D}|\Lambda) - \lambda G(E_{\Lambda}[\mathbf{f}(w,s)]) \tag{7}$$

where λ is a penalized parameter which controls the relative influence of the prior knowledge. This parameter is empirically set to 100 for all the datasets. For brevity, we omit λ in the subsequent derivations. The learning of the LSM model is to maximize the objective function in Equation 7. Exact inference on the LSM is intractable. We use the variational methods to approximate the posterior distribution over the latent variables. The variational distribution which is assumed to be fully factorized is:

$$q(\mathbf{s},\boldsymbol{\theta},\boldsymbol{\varphi}|\Omega) = \prod_{s=1}^{S} q(\varphi_s|\tilde{\beta}_s) \prod_{d=1}^{M} q(\theta_d|\tilde{\alpha}_d) \prod_{t=1}^{N} q(s_{dt}|\tilde{\gamma}_{dt})$$

where $\Omega = \{\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}\}$ are free variational parameters, $\theta \sim$ Dirichlet$(\tilde{\alpha})$, $\varphi \sim$ Dirichlet$(\tilde{\beta})$, and $s_{dt} \sim$ Multinomial$(\tilde{\gamma})$.

We can bound the objective function in Equation 7 in the following way.

$$\mathcal{O}(\mathcal{D}|\Lambda) \geq E_q[\log P(\mathbf{w}, \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\Lambda) - G(E_\Lambda[\mathbf{f}(\mathbf{w}, s)])] - E_q[\log q(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\varphi})] \quad (8)$$

By letting $L(\Omega; \Lambda)$ denote the RHS of the above equation, we have:

$$\mathcal{O}(\mathcal{D}|\Lambda) = L(\Omega; \Lambda) + KL(q(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\Omega)||P(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\Lambda))$$

By maximizing the lower bound $L(\Omega; \Lambda)$ with respect to Ω is the same as minimizing the KL distance between the variational posterior probability and the true posterior probability.

Expanding the lower bound by using the factorizations of P and q, we have:

$$\begin{aligned} L(\Omega; \Lambda) = & E_q[\log P(\boldsymbol{\theta}|\alpha)] + E_q[\log P(\boldsymbol{\varphi}|\beta)] + E_q[\log P(\mathbf{s}|\boldsymbol{\theta})] + E_q[\log P(\mathbf{w}|\mathbf{s}, \boldsymbol{\varphi})] \\ & - E_q[\log q(\boldsymbol{\varphi})] - E_q[\log q(\boldsymbol{\theta})] - E_q[\log q(\mathbf{s})] - E_q[G(E_\Lambda[\mathbf{f}(w, s)])] \quad (9) \end{aligned}$$

The first seven terms are the same as in the LDA model. We show how to compute the last term in the above equation. For a sentiment label j

$$\begin{aligned} E_q[G(E_\Lambda[\mathbf{f}(w, j)])] &= E_q[\sum_w \hat{f}_{jw} \log \frac{\hat{f}_{jw}}{E_\Lambda[\mathbf{f}_{jw}(w, j)]}] \\ &\leq \sum_w \hat{f}_{jw}(\log \hat{f}_{jw} - E_q[\sum_{d=1}^{M}\sum_{t=1}^{N_d} \log(P(w_{d,t}|s_{d,t}; \Lambda)\delta(s_{d,t} = j))]) \\ &= \sum_w \hat{f}_{jw}(\log \hat{f}_{jw} - \sum_{d=1}^{M}\sum_{t=1}^{N_d} \tilde{\gamma}_{d,t,s}\delta(s_{d,t} = j)(\Psi(\tilde{\beta}_{j,w}) - \Psi(\sum_{r=1}^{V} \tilde{\beta}_{j,r}))) \end{aligned}$$

We then employ a variational expectation-maximization (EM) algorithm to estimate the variational parameters Ω and the model parameters Λ.

– (E-step): For each word, optimize values for the variational parameters $\Omega = \{\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}\}$. The update rules are

$$\tilde{\alpha}_{d,s} = \alpha + \sum_{t=1}^{N_d} \tilde{\gamma}_{d,t,s} \quad (10)$$

$$\tilde{\beta}_{s,v} = \beta + \sum_{d=1}^{M}\sum_{t=1}^{N_d} \delta(w_{d,t} = v)\tilde{\gamma}_{d,t,s} \quad (11)$$

$$\tilde{\gamma}_{d,t,s} = \begin{cases} \exp(\Psi(\tilde{\alpha}_{d,s}) + (1 + \hat{f}_{s,w_{d,t}})(\Psi(\tilde{\beta}_{s,w_{d,t}}) - \Psi(\sum_v \tilde{\beta}_{s,v}))) & \text{for labeled features} \\ \exp(\Psi(\tilde{\alpha}_{d,s}) + \Psi(\tilde{\beta}_{s,w_{d,t}}) - \Psi(\sum_v \tilde{\beta}_{s,v})) & \text{otherwise} \end{cases} \quad (12)$$

– (M-step): To estimate the model parameters, we maximize the lower bound on the log likelihood with respect to the parameters $\Lambda = \{\alpha, \beta\}$. There are no closed form solution for α and β and an iterative searching algorithm is used to find the maximal values.

4 Experimental Setup

We conducted experiments on the four corpora[3] which were derived from product reviews harvested from the website IT1681[4] with each corresponding to different types of product reviews including mobile phones, digital cameras, MP3 players, and monitors. All the reviews were tagged by their authors as either positive or negative overall. The total number of review documents is 5484. Chinese word segmentation was performed on the four corpora using the conditional random fields based Chinese Word Segmenter[5].

5 Experimental Results

This section presents the experimental results obtained using the LSM model with translated English lexicons tested on the Chinese product review corpora. The results are averaged over five runs using different random initialization.

5.1 Results with Different Sentiment Lexicons

We explored three widely used English sentiment lexicons in our experiments, namely the MPQA subjectivity lexicon, the appraisal lexicon[6], and the SentiWordNet[7] [5]. For all these lexicons, we only extracted words bearing positive or negative polarities and discarded words bearing neutral polarity. For SentiWordNet, as it consists of words marked with positive and negative orientation scores ranging from 0 to 1, we extracted a subset of 8,780 opinionated words, by selecting those whose orientation strength is above a threshold of 0.6.

We used Google translator toolkit[8] to translate these three English lexicons into Chinese. After translation, duplicate entries, words that failed to translate, and words with contradictory polarities were removed. For comparison, we also tested a Chinese sentiment lexicon, NTU Sentiment Dictionary (NTUSD)[9] which was automatically generated by enlarging an initial manually created seed vocabulary by consulting two thesauri, the Chinese Synonym Thesaurus (tong2yi4ci2ci2lin2) and the Academia Sinica Bilingual Ontological WordNet 3.

Table 1 gives the classification accuracy results using the LSM model with prior sentiment label information provided by different sentiment lexicons. As for the individual lexicon, using the MPQA subjectivity lexicon outperforms the others on all the four corpora. In fact, it performs even better than the Chinese sentiment lexicon

[3] http://www.informatics.sussex.ac.uk/users/tz21/dataZH.tar.gz
[4] http://product.it168.com
[5] http://nlp.stanford.edu/software/stanford-chinese-segmenter-2008-05-21.tar.gz
[6] http://lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz
[7] http://sentiwordnet.isti.cnr.it/
[8] http://translate.google.com
[9] http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html

Table 1. Sentiment classification accuracy (%) by LSM with different sentiment lexicon.

Lexicon	*Mobile*	*DigiCam*	*MP3*	*Monitors*	*Average*
(a) MPQA	80.95	78.65	81.85	79.91	80.34
(b) Appraisal	79.76	70.54	75.84	72.89	74.76
(c) SentiWN	76.06	66.90	75.23	70.28	72.12
(d) NTUSD	80.10	74.17	78.41	79.71	78.10
(a)+(b)	77.20	74.13	77.56	76.93	76.46
(a)+(d)	82.03	80.18	81.03	82.40	81.41
(a)+(b)+(d)	79.21	78.91	77.25	80.85	79.06

NTUSD. The above results suggest that in the absence of any Chinese sentiment lexicon, the translated MPQA subjectivity lexicon can be used to provide sentiment prior information to the LSM model for cross-lingual sentiment classification.

We also conducted experiments by enlarging the MPQA subjectivity lexicon through adding unseen lexical terms from the Appraisal lexicon and NTUSD. We found that the enlargement of a sentiment lexicon does not necessarily lead to the improvement of classification accuracy. In particular, adding new lexical terms from the Appraisal lexicon hurts the classification performance. However, adding extra lexical terms from NTUSD gives the best overall classification accuracy with 81.41% being achieved. Thus, in all the subsequent experiments, the sentiment prior knowledge was extracted from the combination of MPQA subjectivity lexicon and NTUSD.

5.2 Comparison with Other Models

We compare our proposed approach with several other methods as described below:

- *Lexicon labeling*. We implemented a baseline model which simply assigns a score +1 and -1 to any matched positive and negative word respectively based on a sentiment lexicon. A review document is then classified as either positive or negative according to the aggregated sentiment score. Thus, in this baseline model, a document is classified as positive if there are more positive words than negative words in the document and vice versa.
- *LDA*. We evaluated sentiment classification performance with the LDA model where the number of topics were set to 3 corresponding to the 3 sentiment labels.
- *LDA init with prior*. The word prior polarity information obtained from a sentiment lexicon is incorporated during the initialization stage of LDA model learning. Each word token in the corpus is compared against the words in a sentiment lexicon. The matched word token get assigned its prior sentiment label. Otherwise, it is assigned with a randomly selected sentiment label.
- *LSM with random init*. The LSM model is trained with random initialization. That is, the word prior sentiment information is only incorporated by modifying the LDA objective function.
- *LSM init with prior*. Similar to *LDA init with prior*, the word prior polarity information is also used to initialize the LSM model.

Table 2. Sentiment classification accuracy (%) using different models.

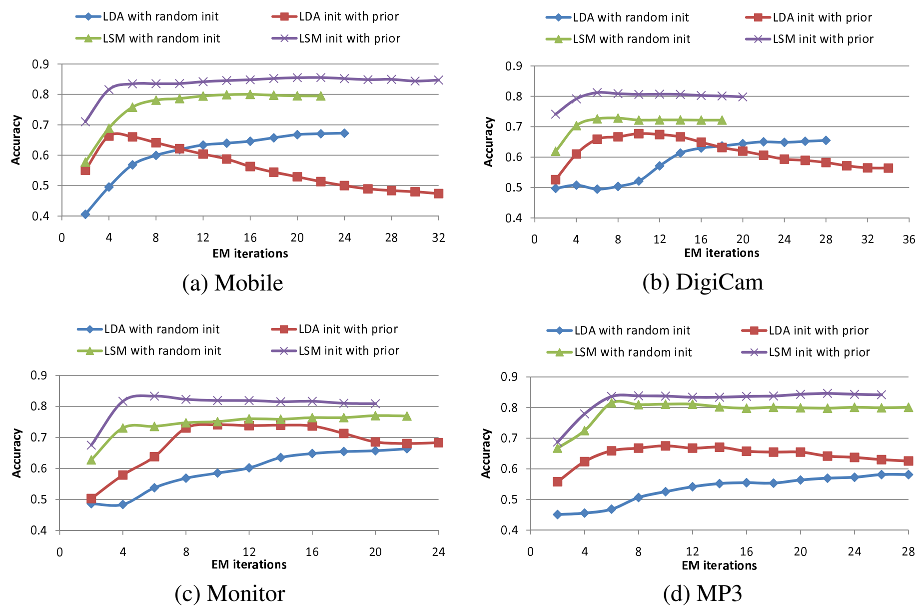
Corpus	*Lexicon Labeling*	*LDA*	*LDA init with prior*	*LSM with random init*	*LSM init with prior*	*Naïve Bayes*	*SVM*
Mobile	68.48	54.88	62.88	74.17	82.03	86.52	84.49
DigiCam	71.70	58.86	62.36	65.91	80.18	82.27	82.04
MP3	70.44	63.11	70.40	74.48	81.03	82.64	79.43
Monitor	71.11	68.82	69.08	81.11	82.40	84.21	83.87
Average	70.43	61.42	66.18	73.92	81.41	84.41	82.46

Table 2 shows the classification accuracy results on the four corpora using different models. It can be observed that *Lexicon labeling* achieves the accuracy in the range of 68-72% with an average of 70.43% obtained over all the four corpora. LDA model without incorporating any sentiment prior information performs quite poorly with its accuracy being only better than random classification. An improvement is observed if the prior sentiment knowledge is incorporated during the initialization stage of LDA model learning. Still, the results are worse than the simple *Lexicon labeling*. For the LSM model, if the prior sentiment knowledge is only used to modify the LDA objective function, it improves upon *LDA init with prior* 4-13%. By additionally incorporating the prior sentiment information into the initialization stage of model learning, LSM outperforms all the other models with the best accuracy of 81.41% being achieved. Compared to *Lexicon labeling*, the improvement obtained by *LSM init with prior* ranges between 11% and 14% and this roughly corresponds to how much the model learned from the data. We can thus speculate that LSM is indeed able to learn the sentiment-word distributions from data.

For comparison purposes, we list the 10-fold cross validation results obtained using the supervised classifiers, Naïve Bayes and SVMs, trained on the labeled corpora [18]. It can be observed that the weakly-supervised LSM model performs comparably to SVMs and is only slightly worse than Naïve Bayes on the Chinese product review corpora despite using no labeled documents.

5.3 Impact of Prior Information

To further investigate the impact of the prior information on model learning, we plot the classification accuracy versus the EM iterations. Figure 2 shows the results on the four corpora. We notice that accuracies of all the other three models except *LDA init with prior* improve with the increasing number of EM iterations. Both *LSM with random init* and *LSM init with prior* converge quite fast, with the best classification accuracy results being achieved after six iterations. *LDA with random init* takes longer time to converge that it gives the best result after 20-26 iterations. The accuracy of *LDA init with prior* reaches the peak after 4 to 8 iterations and then gradually drops. This is more noticeable on Mobile and DigiCam where the accuracy drops about 17% and 11% respectively from the peak. This shows that incorporating prior information only at the initialization stage is not effective since the model is likely to migrate from the initialization state with the increasing number of iterations and thus results in degraded performance. Using the word prior sentiment knowledge to modify the LDA objective function, *LSM with random init* improves over the best results of *LDA init with prior* by 3-14% and it gives more stable results. By incorporating the prior information in both the initialization stage and objective function modification, *LSM init with prior* gives the best results among all the four models.



(a) Mobile (b) DigiCam

(c) Monitor (d) MP3

Fig. 2. Classification accuracy vs EM iterations.

5.4 Domain-Specific Polarity-Bearing Words

While a generic sentiment lexicon provides useful prior knowledge for sentiment analysis, the contextual polarity of a word may be quite different from its prior polarity. Also, the same word might have different polarity in different domain. For example, the word "*compact*" might express positive polarity when used to describe a digital camera, but it could have negative orientation if it is used to describe a hotel room. Thus, it is worth to automatically distinguish between prior and contextual polarity. Our proposed LSM model is able to extract domain-specific polarity-bearing words. Table 3 lists some of the polarity words identified by the LSM model which are not found in the original sentiment lexicons. We can see that LSM is indeed able to recognize domain-specific positive or negative words, for example, 人性化 (user-friendly) for mobile phones, 小巧 (compact) for digital cameras, 金属 (metallic) and 杂音 (noise) for MP3, 清晰 (in focus) and 漏光 (leakage of light) for monitors.

The iterative approach proposed in [18] can also automatically acquire polarity words from data. However, it appears that only positive words were identified by their

Table 3. Extracted example polarity words by LSM.

Corpus		*Extracted Polarity Words*
Mobile	Pos	不错 (not bad;pretty good), 人性化 (user-friendly), 时尚 (fashioable), 好用 (easy to use), 小巧 (compact), 舒服 (comfortable), 薄 (thin;light) 蓝牙 (blue-tooth), 强 (strong;strength), 容易 (easy)
	Neg	坏 (bad), 差 (poor), 死机 (clash), 慢 (slow), 没 (no;not), 难 (difficult;hard), 少 (less), 修 (repair)
DigiCam	Pos	简单 (simple), 防抖 (shake reduction), 优点 (advantage), 小巧 (compact), 时尚 (fashionable), 强 (strong;strength), 长焦 (telephoto), 动态 (dynamic), 全 (comprehensive), 专业 (professional)
	Neg	后悔 (regret), 坏 (bad), 差 (poor), 退货 (return;refund), 慢 (slow), 暗 (dark), 贵 (expensive), 难 (difficult;hard), 耗电 (consume much electricity), 塑料 (plastic), 修 (repair)
MP3	Pos	出色 (outstanding), 小巧 (compact), 齐全 (comprehensive) 简单 (simple), 强 (strong;strength), 美观 (beautiful), 质感 (textual), 金属 (metallic), 不错 (not bad;pretty good)
	Neg	杂音 (noise), 费电 (consume much electricity), 差 (poor), 坏 (bad), 短 (short), 贵 (expensive), 次 (substandard), 死机 (crash), 没 (no), 但是 (but)
Monitors	Pos	专业 (professional), 清晰 (in focus), 时尚 (fashionable), 简洁 (concise), 节能 (energy efficient), 纯平 (flat screen), 不错 (not bad;pretty good), 舒服 (comfortable), 显亮 (looks bright), 锐利 (sharp)
	Neg	变形 (deformation), 模糊 (blurred), 严重 (serious;severe), 失真 (distortion), 偏色 (color cast bad), 坏 (bad), 差 (poor), 漏光 (leakage of light), 黑屏 (black screen), 暗 (dark), 抖动 (jitter)

approach. Our proposed LSM model can extract both positive and negative words and most of them are highly domain-salient as can be seen from Table 3.

6 Conclusions

This paper has proposed the latent sentiment model (LSM) for weakly-supervised cross-lingual sentiment classification. A mechanism has been introduced to incorporate prior information about polarity words from sentiment lexicons where preferences on expectations of sentiment labels of those lexicon words are expressed using generalized expectation criteria. Experimental results of sentiment classification on Chinese product reviews show that in the absence of a language-specific sentiment lexicon, the translated English lexicons can still produce satisfactory results with the sentiment classification accuracy of 80.34% being achieved averaging over four different types of product reviews. Compared to the existing approaches to cross-lingual sentiment classification which either rely on labeled corpora for classifier learning or iterative training for performance gains, the proposed approach is simple and readily to be used for online and real-time sentiment classification from the Web.

One issue relating to the proposed approach is that it still depends on the quality of machine translation and the performance of sentiment classification is thus affected by the language gap between the source and target language. A possible way to alleviate this problem is to construct a language-specific sentiment lexicon automatically from data and use it as the prior information source to be incorporated into LSM model learning.

References

1. Banea, C., Mihalcea, R., Wiebe, J., Hassan, S.: Multilingual subjectivity analysis using machine translation. In: Proceedings of the EMNLP. pp. 127–135 (2008)
2. Bautin, M., Vijayarenu, L., Skiena, S.: International sentiment analysis for news and blogs. In: Proceedings of the ICWSM (2008)
3. Blei, D., McAuliffe, J.: Supervised topic models. Advances in Neural Information Processing Systems 20, 121–128 (2008)
4. Druck, G., Mann, G., McCallum, A.: Learning from labeled features using generalized expectation criteria. In: SIGIR. pp. 595–602 (2008)
5. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC. vol. 6 (2006)
6. Lacoste-Julien, S., Sha, F., Jordan, M.: DiscLDA: Discriminative learning for dimensionality reduction and classification. In: NIPS (2008)
7. Lin, C., He, Y., Everson, R.: A Comparative Study of Bayesian Models for Unsupervised Sentiment Detection. In: CoNLL (2010)
8. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: CIKM (2009)
9. McCallum, A., Mann, G., Druck, G.: Generalized expectation criteria. Tech. Rep. 2007-60, University of Massachusetts Amherst (2007)
10. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the ACL. pp. 976–983 (2007)
11. Mimno, D., McCallum, A.: Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In: Proceedings of the UAI (2008)
12. Qiu, L., Zhang, W., Hu, C., Zhao, K.: Selc: a self-supervised model for sentiment classification. In: Proceeding of the CIKM. pp. 929–936 (2009)
13. Ramage, D., Hall, D., Nallapati, R., Manning, C.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP. pp. 248–256 (2009)
14. Schapire, R., Rochery, M., Rahim, M., Gupta, N.: Incorporating prior knowledge into boosting. In: ICML. pp. 538–545 (2002)
15. Tan, S., Wang, Y., Cheng, X.: Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In: Proceedings of the SIGIR. pp. 743–744 (2008)
16. Wan, X.: Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the EMNLP. pp. 553–561 (2008)
17. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the ACL and the AFNLP. pp. 235–243 (2009)
18. Zagibalov, T., Carroll, J.: Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the COLING. pp. 1073–1080 (2008)
19. Zagibalov, T., Carroll, J.: Unsupervised classification of sentiment and objectivity in chinese text. In: Proceedings of the IJCNLP. pp. 304–311 (2008)