Transfer Learning for Multiple-Domain Sentiment Analysis — Identifying Domain Dependent/Independent Word Polarity

Yasuhisa Yoshida[†] Tsutomu Hirao[††] Tomoharu Iwata[††] Masaaki Nagata[††] Yuji Matsumoto[†]

†Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{yasuhisa-y, matsu}@is.naist.jp

†† NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{hirao.tsutomu, iwata.tomoharu, nagata.masaaki}@lab.ntt.co.jp

Abstract

Sentiment analysis is the task of determining the attitude (positive or negative) of documents. While the polarity of words in the documents is informative for this task, polarity of some words cannot be determined without domain knowledge. Detecting word polarity thus poses a challenge for multiple-domain sentiment analysis. Previous approaches tackle this problem with transfer learning techniques, but they cannot handle multiple source domains and multiple target domains. This paper proposes a novel Bayesian probabilistic model to handle multiple source and multiple target domains. In this model, each word is associated with three factors: Domain label, domain dependence/independence and word polarity. We derive an efficient algorithm using Gibbs sampling for inferring the parameters of the model, from both labeled and unlabeled texts. Using real data, we demonstrate the effectiveness of our model in a document polarity classification task compared with a method not considering the differences between domains. Moreover our method can also tell whether each word's polarity is domain-dependent or domain-independent. This feature allows us to construct a word polarity dictionary for each domain.

1 Introduction

Sentiment analysis is recognized as an important research area in recent years, fueled by the rapid increase of opinion information available from the Internet. A major task of sentiment analysis is that of classifying documents by their polarity; i.e., whether a document is written with a positive attitude or a negative attitude towards the central topic of the document.

In sentiment analysis, a word's polarity is often used as a feature for machine learning. However, the polarity of some words cannot be determined without domain knowledge. Take the word 'long', for example. In the Camera domain, 'long' may have a positive polarity, as in 'the battery life of Camera X is long'. In the Computer domain, on the other hand, it can be negatively polarized, as in 'Program X takes a long time to complete'. Thus it is not easy to construct word polarity dictionaries for all the domains of interest.

A natural question is whether we can reuse the learned result of a certain domain in another. This is indeed a motivation behind transfer learning (also known as domain adaptation or covariate shift). Transfer learning utilizes the result learned in a certain domain (source domain) to solve a similar problem in another new domain (target domain) (Pan and Yang 2008). Research in transfer learning often distinguishes domain-dependent features from independent ones (Blitzer, McDonald, and Pereira 2006), (Daumé III and Marcu 2006).

Previous approaches to sentiment analysis based on transfer learning have a critical problem in that they only work well on a single source domain and a single target domain. In real text data (for example, Multi-Domain Sentiment Dataset (Blitzer, Dredze, and Pereira 2007)) there is a wide variety of domains. Therefore we need a novel method for transfer learning to handle multiple source domains and multiple target domains.

This paper proposes a novel model to handle multiple source domains and multiple target domains. Our model describes how each word is generated in the manner of Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), but takes into consideration three factors in generating a word: Domain label, domain dependence/independence and word polarity. Figure 1 shows the concept of our proposed model in brief.

Our contributions in this paper are as follows.

1. We propose a novel probabilistic generative model of words with a domain label, domain dependence/independence and a word polarity. These factors enable us to handle multiple source domains and multiple target domains. This feature is important because data used in sentiment analysis range over various domains.
2. Our model can tell whether a certain word's polarity is domain-dependent or domain-independent.
3. We construct a sampling method using Gibbs Sampling to calculate the posterior distribution effectively.
4. We apply the proposed model to real product review data (Blitzer, Dredze, and Pereira 2007), and demonstrate its effectiveness compared with the baseline model that does not take into account the domain-dependence/independence of words.

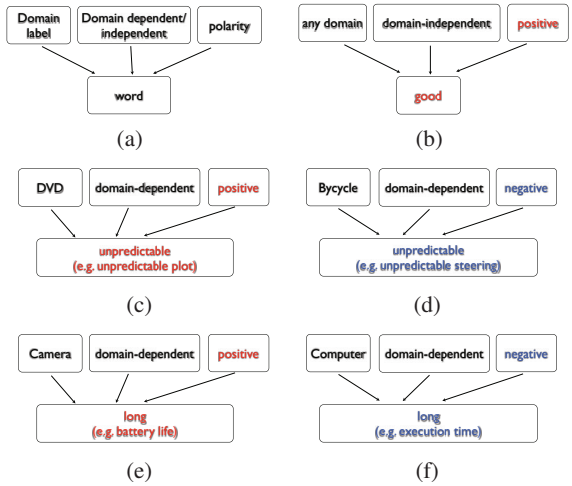


Figure 1: Conceptual diagram of our proposed method. (a) Generative process of a word. A word is generated by a word polarity, a domain label and a domain dependent/independent variable. (b)-(f) Examples of the generative process for five words (words with red-colored mean positive and words with blue-colored mean negative). Previous transfer learning approaches cannot handle multiple domains without constructing multiple models for each pair of domains. Our model can handle them with a single model.

2 Proposed Method

Previous transfer learning approaches are applicable only when there is a single source domain and a single target domain. As we mentioned above, it is important to construct a model to handle multiple source domains and multiple target domains taking the differences of domains into consideration because it is common in sentiment analysis that there are various kind of domains. Our model is designed with multiple source and target domains in mind from the outset.

In this section, we describe the details of our proposed method. We summarize the meaning of symbols in Table 1.

2.1 Model description

In our proposed method, a word polarity l_n and a domain dependence/independence variable z_n are modeled as hidden variables for each word w_n. With l_n and z_n, we have $|Z| \times |S|$ combinations for each word w_n where $Z = \{\text{dependent}, \text{independent}\}$, $S = \{\text{positive}, \text{negative}, \text{neutral}\}$. As we also consider domains, a word w_n can take $|S| \times |Z| \times |F|$ different states. However, domain-independent word can be enclosed with positive, negative, and neutral. After all, the total number of word states is $|S| \times (|F|+1)$. As we can regard domain label f as observed, we can treat the difference between source and target domains. Note that it is easy to extend the proposed model to the cases where there is partially class labeled data in the target domains or there is not partially class labeled data in the source domains.

Next, we explain how the polarity of d-th document y_d is generated. In this model, $P(y_d = m)$ is determined by the proportion to the number of polarity m in $\{l_1, \cdots, l_{N_d}\}$ in the d-th document. To determine $P(y_d)$, more sophisticated methods could be applied such as the Maximum Entropy method as it can use richer features including POS tags (Daumé III and Marcu 2006). To keep the model simple, we did not employ such a method.

Table 1: Notation

Symbol	Description
D, D_s, D_t	number of documents, number of documents in the source domain, number of documents in the target domain ($D = D_s + D_t$)
F	set of domains
S	set of word polarity (in this paper, we fixed this as $\{\text{positive}, \text{negative}, \text{neutral}\}$)
V	number of unique words
N_d	number of words in the d-th document
$w_{d,n}$	n-th word in the d-th document
$y_d, y_d^{(s)}, y_d^{(t)}$	polarity of the d-th document, polarity of the d-th document in the source domain, polarity of the d-th domain in the target domain (0 means negative, and 1 means positive)
$z_{d,n}$	indicator variable for n-th word in the d-th document representing whether it is domain-specific of not
$l_{d,n}$	word polarity of $w_{d,n}$, $l_{d,n} = 0$ if negative, $l_{d,n} = 1$ if positive, $l_{d,n} = 2$ if neutral
f_d	domain of the d-th document

In summary, the proposed model assumes the following generative process,

1. For each sentiment $l \in S$:
   (a) Draw word probability $\phi_{0,l} \sim \text{Dirichlet}(\beta)$
   (b) For each domain $f \in F$:
      i. Draw word probability $\phi_{1,l,f} \sim \text{Dirichlet}(\beta)$
2. For each document $d = 1, \cdots, D$:
   (a) Draw domain dependent probability $\theta_d \sim \text{Beta}(\alpha)$
   (b) Draw word sentiment probability $\psi_d \sim \text{Dirichlet}(\gamma)$
   (c) For each word $w_{d,n}$ in document d
      i. Draw $z_{d,n} \sim \text{Bernoulli}(\theta_d)$
      ii. Draw $l_{d,n} \sim \text{Multinomial}(\psi_d)$
   (d) Draw $w_{d,n} \sim \begin{cases} \text{Multinomial}(\phi_{0,l_{d,n}}) & (\text{if } z_{d,n} = 0) \\ \text{Multinomial}(\phi_{1,l_{d,n},f_d}) & (\text{if } z_{d,n} = 1) \end{cases}$
   (e) Draw $y_d \sim \text{Bernoulli}\left(\frac{N_{d,0}+\eta}{N_{d,0}+N_{d,1}+2\eta}, \frac{N_{d,1}+\eta}{N_{d,0}+N_{d,1}+2\eta}\right)$

where $\phi_{0,l}, \phi_{1,l,f}, \theta_d, \psi_d$ are parameter vectors whose elements are $\phi_{0,l,w} = P(w|z=0,l), \phi_{1,l,f,w} = P(w|z=1,l,f)$. $\theta_{d,z}$ and $\psi_{d,l}$ are $P(z)$ and $P(l)$ in the d-th document. $N_{d,0}$ is the number of negative words in the d-th document, and $N_{d,1}$ is the number of positive words in the d-th document.

The Dirichlet distributions have parameters α, β and γ. We assume they are symmetric. We also show the graphical model of this generative process in Figure 2.1. The joint probability of this model can be written as follows:

$$P(\mathbf{w}, \mathbf{y}, \mathbf{z}, \mathbf{l} | \mathbf{f}, \alpha, \beta, \gamma, \eta) = P(\mathbf{z}|\alpha) P(\mathbf{w}|\mathbf{z}, \mathbf{f}, \mathbf{l}, \beta) P(\mathbf{l}|\gamma) P(\mathbf{y}|\mathbf{l}, \eta), \quad (1)$$

where $\mathbf{w} = \{\{w_{d,n}\}_{n=1}^{N_d}\}_{d=1}^{D}, \mathbf{z} = \{\{z_{d,n}\}_{n=1}^{N_d}\}_{d=1}^{D}, \mathbf{l} = \{\{l_{d,n}\}_{n=1}^{N_d}\}_{d=1}^{D}, \mathbf{f} = \{f_d\}_{d=1}^{D}, \mathbf{y} = \{y_d\}_{d=1}^{D}$. We can

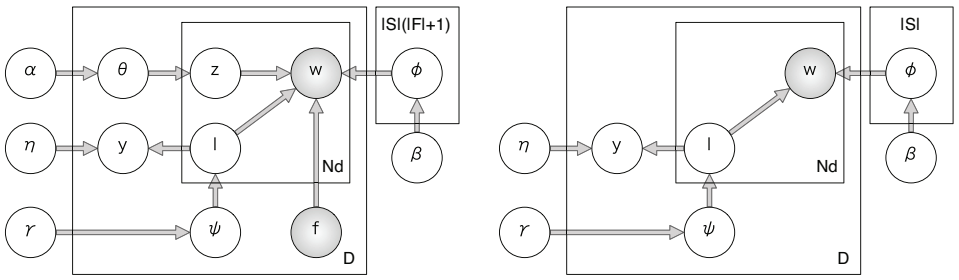


Figure 2: (a) Graphical model of proposed method. (b) Graphical model simplifying proposed method. This model is the baseline in the experiments.

integrate out the multinomial distribution parameters, $\{\phi_{0,l}\}_{l\in S}, \{\phi_{1,l,f}\}_{l\in S, f\in F}, \{\theta_d\}_{d=1}^D, \{\psi_d\}_{d=1}^D$, because we use Dirichlet distributions for their priors, which are conjugate to multinomial distributions. The first term on the right side of (1) is calculated by $P(\mathbf{z}|\alpha) = \prod_{d=1}^D \int P(\mathbf{z}_d|\theta_d)P(\theta_d|\alpha)d\theta_d$, and we have the following equation by integrating out $\{\theta_d\}_{d=1}^D$,

$$P(\mathbf{z}|\alpha) = \left(\frac{\Gamma(|Z|\alpha)}{\Gamma(\alpha)^{|Z|}}\right)^D \prod_{d=1}^D \frac{\prod_{z\in Z}\Gamma(N_{z,d}+\alpha)}{\Gamma(\sum_{z\in Z}(N_{z,d}+\alpha))}, \quad (2)$$

where $N_{z,d}$ is the number of words that are domain dependent/independent in the d-th document. The rest of the joint probabilities on the right side of (1) can be written as follows:

$$P(\mathbf{w}|\mathbf{z},\mathbf{f},\mathbf{l},\beta) = \left(\frac{\Gamma(\beta V)}{\Gamma(\beta)^V}\right)^{(|F|+1)\times|S|} \times \prod_{f\in F}\prod_{s\in S}\frac{\prod_{v=1}^V \Gamma(N_{v,1,s,f}+\beta)}{\Gamma(\sum_{v=1}^V (N_{v,1,s,f}+\beta))} \times \prod_{s\in S}\frac{\prod_{v=1}^V \Gamma(N_{v,0,s}+\beta)}{\Gamma(\sum_{v=1}^V (N_{v,0,s}+\beta))},$$

$$P(\mathbf{l}|\gamma) = \left(\frac{\Gamma(|S|\gamma)}{\Gamma(\gamma)^{|S|}}\right)^D \prod_{d=1}^D \frac{\prod_{s\in S}\Gamma(N_{s,d}+\gamma)}{\Gamma(\sum_{s\in S}(N_{s,d}+\gamma))},$$

where $N_{v,1,s,f}$ is the number of domain-dependent words v with polarity s in domain f, $N_{v,0,s}$ is the number of domain-independent words v with polarity s, $N_{s,d}$ is the number of words with polarity s in d-th document. The fourth term on the right side of (1) is a multinomial distribution,

$$P(\mathbf{y}|\mathbf{l},\eta) = \prod_{d=1}^D \prod_{y=0}^1 \left(\frac{N_{d,y}+\eta}{N_{d,0}+N_{d,1}+2\eta}\right)^{I(y_d=y)}, \quad (3)$$

where $N_{d,y}$ is the number of words with polarity y in the d-th document, $N_{d,0}$ is the number of negative words in the d-th document, $N_{d,1}$ is the number of positive words in the d-th document, and $I(\cdot)$ is an indicator function.

2.2 Inference

The inference of the latent variables $\mathbf{z}$, $\mathbf{l}$ and $\mathbf{y}^{(t)}$ given $\mathbf{w}$ and $\mathbf{f}$ can be effectively computed using collapsed Gibbs sampling (Griffiths and Steyvers 2004). Note that in the source domains, the latent variables are $\mathbf{z}$ and $\mathbf{l}$, but in the target domains, the latent variables are $\mathbf{z}$, $\mathbf{l}$ and $\mathbf{y}^{(t)}$. Sampling formula in the source domains can be written as follows:

$$\begin{aligned} &P(z_n = j, l_n = k|\mathbf{w},\mathbf{f},\mathbf{y}^{(s)},\mathbf{z}_{-n},\mathbf{l}_{-n},\alpha,\beta,\gamma,\eta) \\ &\propto P(w_n|z_n = j, l_n = k, f_d, \mathbf{z}_{-n},\mathbf{l}_{-n},\beta)P(z_n = j|\mathbf{z}_{-n},\alpha) \\ &\times P(y_d^{(s)}|l_n = k,\mathbf{l}_{-n},\eta)P(l_n = k|\mathbf{l}_{-n},\gamma), \end{aligned} \quad (4)$$

where $\mathbf{y}^{(s)} = \{y_d^{(s)}\}_{d=1}^{D_s}$. The elements of the right side of (4) can be written as follows:

$$P(w_n|z_n = 1, l_n = k, f_d, \mathbf{z}_{-n},\mathbf{l}_{-n},\beta) = \frac{\{N_{w_n,1,k,f_d}\}_{-n}+\beta}{\sum_{w_n=1}^V(\{N_{w_n,1,k,f_d}\}_{-n}+\beta)},$$

$$P(w_n|z_n = 0, l_n = k, f_d, \mathbf{z}_{-n},\mathbf{l}_{-n},\beta) = \frac{\{N_{w_n,0,k}\}_{-n}+\beta}{\sum_{w_n=1}^V(\{N_{w_n,0,k}\}_{-n}+\beta)},$$

$$P(z_n = j|\mathbf{z}_{-n},\alpha) = \frac{\{N_{j,d}\}_{-n}+\alpha}{\sum_{j\in Z}(\{N_{j,d}\}_{-t}+\alpha)},$$

$$P(y_d^{(s)}|l_n = k,\mathbf{l}_{-n},\eta) \propto \frac{(\{N_{d,k}\}_{-n}+\eta+1)^{I(y_d^{(s)}=k)}(\{N_{d,1-k}\}_{-n}+\eta)^{I(y_d^{(s)}=1-k)}}{\{N_{d,0}\}_{-n}+\{N_{d,1}\}_{-n}+2\eta+1},$$

$$P(y_d^{(s)}|l_n = 2,\mathbf{l}_{-n},\eta) \propto \prod_{y=0}^1\left(\frac{\{N_{d,y}\}_{-n}+\eta}{\{N_{d,0}\}_{-n}+\{N_{d,1}\}_{-n}+2\eta}\right)^{I(y_d^{(s)}=y)},$$

$$P(l_n = k|\mathbf{l}_{-n},\gamma) = \frac{\{N_{k,d}\}_{-n}+\gamma}{\sum_{k\in S}(\{N_{k,d}\}_{-n}+\gamma)},$$

where $\{\cdot\}_{-n}$ is the number when n-th sample was excluded.

Next, in the target domains, sampling formula for $\mathbf{z}$ and $\mathbf{l}$ is the same as equation (4). In the target domains, we sample $\mathbf{y}^{(t)}$ following the next equation,

$$P(y_d^{(t)} = m|\mathbf{l},\eta) = \frac{N_{d,m}+\eta}{N_{d,0}+N_{d,1}+2\eta}, \quad (5)$$

where $N_{d,m}$ is the number of words with polarity m in the d-th document.

3 Experiments

As mentioned above, our proposed method is a method to handle multiple source domains and multiple target domains. In order to see whether such a characteristic is useful or not in a document polarity classification task, we compare our model shown in Figure 2.1, with the baseline model

which domain dependent/independent variables **z** and domain labels **f** are removed. The resulting graphical model for the baseline is shown in Figure 2.1.

Several ways of changing the number of source/target domains are conceivable. We performed the following two experiments.

- Fixing the number of target domains, increase the number of source domains
- Fixing the number of source domains, increase the number of target domains

The former is reported in Section 3.2, and the latter in Section 3.3. Our method can tell whether a certain word is a domain dependent/independent polarised one in each domain. To show how well our method works, in Section 3.4, we list the most probable domain-dependent and domain-independent polarised words in each domain.

3.1 Dataset and experimental setting

We used 17 domains and 10000 documents from the Multi-Domain Sentiment Dataset (Blitzer, Dredze, and Pereira 2007) for our experiments. In this dataset, each review text is accompanied with five-staged rating. We regard reviews with rating 4 or 5 as positive documents and review with rating 1 or 2 as a negative documents. We adopt POS-filtering, using only adjectives and adverbs. Negated phrases like 'not ...' are regarded as a single word.

Next we describe the experimental setting. We set the number of iteration for the Gibbs sampling to 300, and determined the final documents polarity according to last 50 samples from the right side of (5). Hyper parameters α, β and γ are optimized by Fixed Point Iteration (Minka 2003). For lack of space, we omit the update equations. From preliminary experiments, we fixed the hyper parameter η as 1.5. Note that the performance of documents polarity classification is not very sensitive to the value of η.

And finally, we describe the method of evaluation. We evaluate the performance of documents polarity classification using F-value, harmonic mean of Precision and Recall. We trained our model with both training data and test data because our proposed method is so-called Transductive Learning.

3.2 Increasing the number of source domains fixing target domains

In our experiment, classification accuracy varies by combinations between source domains and target domains. To take this variance into consideration, we repeated the following procedure 10 times and took the average F-value.

Step 1 Select 3 domains for target domains randomly

Step 2 Select N domains for source domains from the remaining ones randomly

Step 3 Evaluate using F-value at the target domains

Figure 3.3 shows the results of the above experiments with the number of N of source domains varied from 1 to 14. As the number of source domains increases, the number of samples also increases. So F-value increases constantly both in the proposed method and the baseline. The F-value of the proposed method is consistently higher than that of the baseline. This result shows the effectiveness of the domain dependent/independent variables incorporated in the proposed model for each words, which are not present in the baseline model.

3.3 Increasing the number of target domains fixing source domains

In this section, we evaluate the performance of the proposed method and the baseline when the source domains is fixed and the number of target domains is varied. We repeated 10 times the same procedure (Step 1-3) of Section 3.2, except that the three domains chosen in Step 1 make the source domains, and the N domains chosen in the Step 2 constitute the target. Figure 3.3 shows the results as N is increased from 1 to 14. F-value decreased in both proposed method and baseline as the number of target domain is increased. The same phenomenon was reported in semi-supervised leaning. Nigam et al. introduced a weight parameter λ for unlabeled data to reduce the effect of unlabeled data (Nigam et al. 2000). However, in the proposed method, the decrease of F-value is more benign compared to the baseline in Figure 3.3.

3.4 Domain dependent/independent words

Our method can tell whether a given word is dependent or independent for each domain. We list the most probable 30 words using $P(w|l,z,f) = \phi_{z,l,f}$ which is the probability that a certain word w is generated given polarity l and domain dependent/independent z in domain f. Tables 2, 3 and 4 shows the domain dependent/independent words for the source domains of 'Books', 'DVD', and 'Electronics' and the target domain of 'Kitchen'. The results look reasonable, for words such as 'great' and 'bad' are determined as domain-independent, and words like 'comfortable', 'responsive', 'useless', 'functionally' are deemed as domain-dependent.

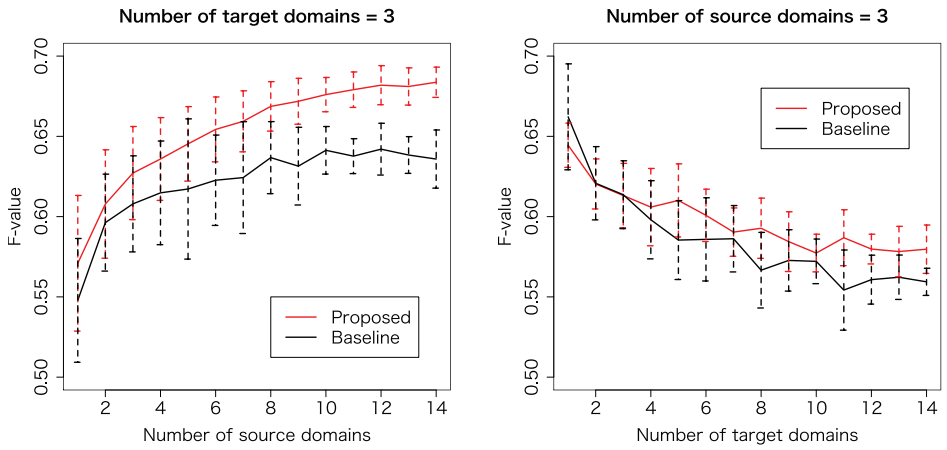
Table 2: Domain-independent words

Polarity	Words
Positive	great good best excellent worth certainly easily happy particularly quickly deep quick not_really professional fantastic incredible solid effective beautifully potential
Negative	bad instead actually wrong unfortunately completely poor worst second short nearly extremely possible worse not_good actual fairly just_not disappointed entirely
Neutral	quite long right away old probably pretty simply big large amazing white free apparently huge exactly forward open normal older

4 Related Work

4.1 Sentiment Analysis

Two types of sentiment analysis tasks have been addressed in the past: Document classification and extraction of features relevant to polarities. The former tries to classify documents according to their semantic orientation such as positive or negative (Pang, Lee, and Vaithyanathan 2002). The



(a) (b)

Figure 3: Average (solid line) and standard error (dotted line) of F-value 10 times experiments (a) F-value in 3 target domains varying the number of source domains. (b) F-value varying the number of target domains with fixed the number of source domains as 3.

Table 3: Domain-dependent words (in Electronics domain)

Polarity	Words
Positive	easy sound small remote fast comfortable perfectly cool external cheaper decent light pleased green crisp compatible higher optical sleek rear
Negative	new cheap useless slow defective directly newer static blank quiet flat uncomfortable convenient currently daily glad flimsy verbatim soft tiny lighter
Neutral	little nice digitallonger low loud inexpensive video audio not_sure properly multiple faulty bulky stylishjust_fine protective manually bright double prior outer

Table 4: Domain-dependent words (in Kitchen domain)

Polarity	Words
Positive	stainless sturdy fresh healthy unique warm spreadable liquid handy serrated largest golly not_big functionally wider ceramic 5-year extendable frozen not_constant
Negative	evenly non-stick durable cuban heavier hefty shiny clean-up not_old broken confident transparent versatile not_short not_probably tinny suspicious appropriately cramped grilled
Neutral	nicely global coated taller dough fluffy thicker distinctive vibrant abrasive visible charcoal vertically bit toxic dear level 5-star in-built playful

latter, on the other hand, focuses on extracting words concerning the polarity of document (Takamura, Inui, and Okumura 2005). Our model can contribute to both of them.

In the task of classifying a document polarity, there is a lot of research using aspects (also known as attributes) and topics. (Zhao et al. 2010) and (Brody and Elhadad 2010) modeled documents by using aspects such as sound quality, battery life, appearance in the Mp3 players domain. Some researchers focus on topics to identify document polarity (Lin, He, and Everson 2010), (Titov and McDonald 2008), (Mei et al. 2007). Lin et al. proposed Joint Sentiment-Topic Model (JST) incorporating both topics and polarity of words in the document (Lin, He, and Everson 2010).

JST, however, does not consider the differences of domains (domain labels)[1]. Our proposed method also represents word polarity using a generative model like JST, but it can also treat domain labels as observed variables. This feature makes our proposed method a more suitable model for transfer learning.

Takamura et al. modeled the polarity of words in a document which provides informative features for detecting document polarity using a spin model (Takamura, Inui, and Okumura 2005). Kobayashi et al. extracted aspect-evaluation relations and aspect-of relations (Kobayashi, Inui, and Matsumoto 2007). Our method contributes to this line of research in that our proposed model can simultaneously construct a domain dependent word polarity dictionary for each domain and a domain independent word polarity dictionary. The work of Takamura et al. is similar to ours in that both can extract words relevant to polarities, but they do not consider the word polarity to be domain-dependent, and thus is not adequate for multi-domain sentiment analysis.

4.2 Transfer Learning

Transfer learning has been studied in various fields including natural language processing (Daumé III and Marcu 2006), automated planning (Li, Yang, and Xue 2009) and collaborative filtering (Zhuo et al. 2008).

Blitzer et al. proposed a framework called Structural Correspondence Learning (SCL) and utilized it for transfer learning (Blitzer, McDonald, and Pereira 2006). SCL is based on the work of Ando et al. (Ando and Zhang 2005). SCL identifies ‘pivot features‘, which are the features having high mutual information with polarity labels, and solve

[1] In this paper, ‘Topic‘ means the hidden variable corresponding to word meaning, and ‘Domain‘ means the observed variable like category.

the auxiliary binary classification problems.

SCL connects pivot features with domain-specific words for each domain in order to guide transfer learning. For example, let source domain be micro phone and target domain be computer. 'good-quality reception' and 'fast dual-core' are the positive words in each domain, but each of these words appears only in each domain respectively. Transfer learning fails if we only use such words. The main idea of SCL is that these words often occur with the general polarized words like 'excellent'. In SCL, however, pivot features are only defined as the words that have high mutual information with the class labels in the source domain, and in this definition, the words with high correlation with the class labels in the target domain can be pivot features. Therefore, we modeled the domain-independent word polarity similar to the approach of Lin et al. (Lin, He, and Everson 2010).

(Daumé III and Marcu 2006) proposed Maximum Entropy Genre Adaptation Model (MEGA) model motivated by the fact that the distribution of test data is not identical to that of training data in many applications. MEGA model is a simple mixture model with a hidden variable indicating whether the data is drawn from the in-domain distribution, the out-of-domain distributions, or the general-domain distribution.

5 Conclusion and Future Work

In this paper, we proposed a probabilistic generative model of a word with a domain label, domain dependence/independence and a word polarity, and this can also judge the document polarity which can treat the differences between domains. Parameter values can be learned with Gibbs sampling. We demonstrated that increased data from source domains lead to an improved F-value in target domains. We also found that as the number of target domains increased, F-value decreased. Our model can extract words with domain-dependent polarity, making it possible to create domain-dependent word polarity dictionaries for each domain. For the future work, we will revise our model not to be sensitive to unlabeled data from target domain when the number of samples in target domain increases. Another direction of the future work is to modify a method for determining a document polarity from the current one to a more sophisticated method based on maximum entropy similar to Daume's MEGA model.

Acknowledgements

We thank Masashi Shimbo and Joseph Irwin for their valuable comments. This work was done during the first author was a visiting student at the NTT CS Labs.

References

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* 6.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. EMNLP '06.

Brody, S., and Elhadad, N. 2010. An unsupervised aspect-sentiment model for online reviews. HLT '10.

Daumé III, H., and Marcu, D. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research (JAIR)* 26.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *PNAS* 101(suppl. 1).

Kobayashi, N.; Inui, K.; and Matsumoto, Y. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining.

Li, B.; Yang, Q.; and Xue, X. 2009. Can movies and books collaborate?: cross-domain collaborative filtering for sparsity reduction. In *Proceedings of the 21st international jont conference on Artifical intelligence*. Morgan Kaufmann Publishers Inc.

Lin, C.; He, Y.; and Everson, R. 2010. A comparative study of bayesian models for unsupervised sentiment detection. CoNLL '10.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. WWW '07. ACM.

Minka, T. P. 2003. Estimating a Dirichlet distribution.

Nigam, K.; Mccallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* 39(2).

Pan, S. J., and Yang, Q. 2008. A Survey on Transfer Learning. Technical report.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques.

Takamura, H.; Inui, T.; and Okumura, M. 2005. Extracting semantic orientations of words using spin model. ACL '05.

Titov, I., and McDonald, R. 2008. Modeling online reviews with multi-grain topic models. WWW '08. ACM.

Zhao, W. X.; Jiang, J.; Yan, H.; and Li, X. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. EMNLP '10.

Zhuo, H.; Yang, Q.; Hu, D. H.; and Li, L. 2008. Transferring knowledge from another domain for learning action models. PRICAI '08.