Clustering Aspect-related Phrases by Leveraging Sentiment Distribution Consistency

Li Zhao, Minlie Huang, Haiqiang Chen*, Junjun Cheng*, Xiaoyan Zhu
State Key Laboratory of Intelligent Technology and Systems
National Laboratory for Information Science and Technology
Dept. of Computer Science and Technology, Tsinghua University, Beijing, PR China
*China Information Technology Security Evaluation Center
zhaoli19881113@126.com aihuang@tsinghua.edu.cn

Abstract

Clustering aspect-related phrases in terms of product's property is a precursor process to aspect-level sentiment analysis which is a central task in sentiment analysis. Most of existing methods for addressing this problem are context-based models which assume that domain synonymous phrases share similar co-occurrence contexts. In this paper, we explore a novel idea, *sentiment distribution consistency*, which states that different phrases (e.g. "price", "money", "worth", and "cost") of the same aspect tend to have consistent sentiment distribution. Through formalizing *sentiment distribution consistency* as soft constraint, we propose a novel unsupervised model in the framework of Posterior Regularization (PR) to cluster aspect-related phrases. Experiments demonstrate that our approach outperforms baselines remarkably.

1 Introduction

Aspect-level sentiment analysis has become a central task in sentiment analysis because it can aggregate various opinions according to a product's properties, and provide much detailed, complete, and in-depth summaries of a large number of reviews. Aspect finding and clustering, a precursor process of aspect-level sentiment analysis, has attracted more and more attentions (Mukherjee and Liu, 2012; Chen et al., 2013; Zhai et al., 2011a; Zhai et al., 2010).

Aspect finding and clustering has never been a trivial task. People often use different words or phrases to refer to the same product property (also called *product aspect or feature* in the literature). Some terms are lexically dissimilar while semantically close, which makes the task more challenging. For example, "price", "money" , "worth" and "cost" all refer to the aspect "price" in reviews. In order to present aspect-specific summaries of opinions, we first of all, have to cluster different aspect-related phrases. It is expensive and time-consuming to manually group hundreds of aspect-related phrases. In this paper, we assume that the aspect phrases have been extracted in advance and we keep focused on clustering domain synonymous aspect-related phrases.

Existing studies addressing this problem are mainly based on the assumption that different phrases of the same aspect should have similar co-occurrence contexts. In addition to the traditional assumption, we develop a new angle to address the problem, which is based on *sentiment distribution consistency* assumption that different phrases of the same aspect should have consistent sentiment distribution, which will be detailed soon later.

Pros:	***LCD***, *nice* ***touch screen***, *longer* **battery life**
Cons:	*Horrible* ***picture quality***
Review:	The ***touch screen*** was the selling feature for me. The ***LCD touch screen*** is nice and large. This camera also has very impressive ***battery life***. However the ***picture quality*** is very grainy.

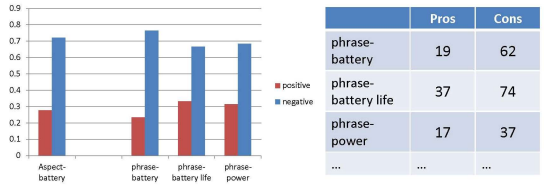
Figure 1: A semi-structured Review.

This new angle is inspired by this simple observation (as illustrated in Fig. 1): two phrases within the same cluster are not likely to be simultaneously placed in Pros and Cons of the same review. A straightforward way to use this information is to formulate cannot-link knowledge in clustering algorithms (Chen et al., 2013; Zhai et al., 2011b). However, we have a particularly different manner to leverage the knowledge.

Due to the availability of large-scale semi-structured customer reviews (as exemplified in Fig. 1) that are supported by many web sites, we can easily get the estimation of sentiment distribution for each aspect phrase by simply counting how many times a phrase appears in *Pros* and

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1614–1623,
October 25-29, 2014, Doha, Qatar.

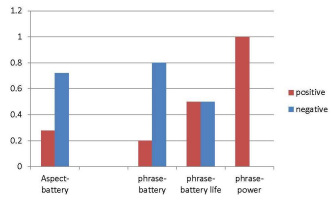
Cons respectively. As illustrated in Fig. 2, we can see that the estimated sentiment distribution of a phrase is close to that of its aspect. The above observation suggests the *sentiment distribution consistency* assumption: **different phrases of the same aspect tend to have the same sentiment distribution, or to have statistically close distributions**. This assumption is also verified by our data: for most (above 91.3%) phrase with relatively reliable estimation (whose occurrence $\geq$50), the KL-divergence between the sentiment distribution of a phrase and that of its corresponding aspect is less than 0.05.



	Pros	Cons
phrase-battery	19	62
phrase-battery life	37	74
phrase-power	17	37
...	...	...

Figure 2: The sentiment distribution of aspect "battery" and its related-phrases on *nokia 5130* with a large amount of reviews.

It is worth noting that, the sentiment distribution of a phrase can be estimated accurately only when we obtain a sufficient number of reviews. When the number of reviews is limited, however, the estimated sentiment distribution for each phrase is unreliable (as shown in Fig. 3). A key issue, arisen here, is how to formulate this assumption in a statistically robust manner. The proposed model should be robust when only a limited number of reviews are available.



	Pros	Cons
phrase-battery	1	4
phrase-battery life	2	2
phrase-power	1	0
...	...	...

Figure 3: The sentiment distribution of aspect "battery" and its related-phrases on *nokia 3110c* with a small mumber of reviews.

To deal with this issue, we model *sentiment distribution consistency* as soft constraint, integrated into a probabilistic model that maximizes the data likelihood. We design the constraint to work in the following way: when we have sufficient observations, the constraint becomes tighter, which plays a more important role in the learning process; when we have limited observations, the constraint becomes very loose so that it will have less effect on the model.

In this paper, we propose a novel unsupervised model, Sentiment Distribution Consistency Regularized Multinomial Naive Bayes (SDC-MNB). The context part is modeled by Multinomial Naive Bayes in which aspect is treated as latent variable, and *Sentiment distribution consistency* is encoded as soft constraint within the framework of Posterior Regularization (PR) (Graca et al., 2008). The main contributions of this paper are summarized as follows:

- We study the problem of clustering phrases by integrating both context information and sentiment distribution of aspect-related phrases.
- We explore a novel concept, *sentiment distribution consistency(SDC)*, and model it as soft constraint to guide the clustering process.
- Experiments show that our model outperforms the state-of-art approaches for aspect clustering.

The rest of this paper is organized as follows. We introduce the SDC-MNB model in Section 2. We present experiment results in Section 3. In Section 4, we survey related work. We summarize the work in Section 5.

2 Sentiment Distribution Consistency Regularized Multinomial Naive Bayes

In this section, we firstly introduce our assumption *sentiment distribution consistency* formally and show how to model the above assumption as soft constraint , which we term SDC-constraint. Secondly, we show how to combine SDC-constraint with the probabilistic context model. Finally, we present the details for context and sentiment extraction.

2.1 Sentiment Distribution Consistency

We define *aspect* as a set of phrases that refer to the same property of a product and each phrase is termed *aspect-related phrase* (or *aspect phrase* in short). For example, the aspect "battery" contains aspect phrases such as "battery", "battery life", "power", and so on.

F	the aspect phrase set
f_j	the j^{th} aspect phrase
y_j	the aspect for aspect phrase f_j
A	the aspect set
a_i	the i^{th} aspect
D	the set of context documents
d_j	the context document of f_j
V	the word vocabulary
w_t	the t^{th} word in vocabulary V
$w_{d_j,k}$	the k^{th} word in d_j
N_{tj}	the number of times word w_t occurs in d_j
P	the product set
p_k	the k^{th} product
u_{ik}	the *sentiment distribution parameter* of aspect a_i on p_k
$\hat{s}_{jk}$	the estimated *sentiment distribution parameter* of phrase f_j on p_k
n_{jk}	the occurrence times of aspect phrase f_j on p_k
$\hat{\sigma}_{jk}$	the sample standard deviation
θ	the model parameters
$p_\theta(a_i\|d_j)$	the posterior distribution of a_i given d_j
$q(y_j = a_i)$	the projected posterior distribution of a_i given d_j

Table 1: Notations

Let us consider the sentiment distribution on a certain aspect a_i. In a large review dataset, aspect a_i could receive many comments from different reviewers. For each comment, we assume that people either praise or complain about the aspect. So each comment on the aspect can be seen as a Bernoulli trial, where the aspect receives positive comments with probability p_{a_i}[1]. We introduce a *random variable* X_{a_i} to denote the sentiment on aspect a_i, where $X_{a_i} = 1$ means that aspect a_i receives positive comments, $X_{a_i} = 0$ means that aspect a_i receives negative comments. Obviously, the sentiment on aspect a_i follows the Bernoulli distribution,

$$Pr(X_{a_i}) = p_{a_i}^{X_{a_i}} * (1 - p_{a_i})^{1 - X_{a_i}}, X_{a_i} \in \{0, 1\}. \quad (1)$$

Or in short,

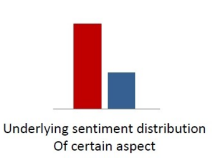
$$X_{a_i} \sim Bernoulli(p_{a_i})$$

Let us see the case for aspect phrase f_j, where $f_j \in$ aspect a_i. Similarly, each comment on an aspect phrase f_j can also be seen as a Bernoulli trial. We introduce a *random variable* X_{f_j} to denote the sentiment on aspect phrase f_j, where $X_{f_j} = 1$ means that aspect f_j receives positive comments, $X_{f_j} = 0$ means that aspect f_j receives negative comments. As just discussed, we assume that each aspect phrase follows the same distribution with the corresponding aspect. This leads to the following formal description:

- ***Sentiment Distribution Consistency*** : The sentiment distribution of *aspect phrase* is the same as that of the corresponding *aspect*. Formally, for all aspect phrase $f_j \in$ aspect a_i, $X_{f_j} \sim Bernoulli(p_{a_i})$.

2.2 Sentiment Distribution Consistency Constraint

Assuming the sentiment distribution of aspect a_i is given in advance, we need to judge whether an aspect phrase f_j belongs to the aspect a_i with limited observations for f_j. Let's consider the example in Fig. 4. For aspect phrase 3, we have no definite answer due to the limited number of observations. For aspect phrase 1, it seems that the sentiment distribution is consistent with that of the left aspect. However, we can not say that the phrase belongs to the aspect because the distribution may be the same for two different aspects. For aspect phrase 2, we are confident that its sentiment distribution is different from that of the left aspect, given sufficient observations.



	pros	cons	Belong to left aspect?
Aspect phrase1	21	10	Reliable estimation, Use to judge. =>maybe
Aspect phrase2	3	100	Reliable estimation, Use to judge. => not likely
Aspect phrase3	1	2	Unreliable estimation =>maybe

Figure 4: Sentiment distribution of an aspect, and observations on aspect phrases.

To be concise, *we judge an aspect phrase doesn't belong to certain aspect only when we are confident that they follow different sentiment distributions*.

Inspired by the intuition, we conduct interval parameter estimation for parameter p_{f_j}(sentiment distribution for phrase f_j) with limited observations, and thus get a confidence interval for p_{f_j}. If p_{a_i}(sentiment distribution for aspect a_i) is not in the confidence interval of p_{f_j}, we then are confident that they follow different distributions. In other words, if aspect phrase $f_j \in$ aspect a_i, we are confident that p_{a_i} is in the confidence interval of p_{f_j}.

More formally, we use u_{ik} to denote the sentiment distribution parameter of aspect a_i on product p_k, and assume that u_{ik} is given in advance.

[1] *positive comment* means that an aspect term is observed in *Pros* of a review.

We want to know whether the sentiment distribution on aspect phrase f_j is the same as that of aspect a_i on product p_k given a limited number of observations (samples). It's straightforward to calculate the confidence interval for parameter s_{jk} in the Bernoulli distribution function. Let the sample mean of n_{jk} samples be $\hat{s}_{jk}$, and the sample standard deviation be $\hat{\sigma}_{jk}$. Since the sample size is small here, we use the Student-t distribution to calculate the confidence interval. According to our assumption, we are confident that u_{ik} is in the confidence interval if $f_j \in a_i$.

$$\hat{s}_{jk} - C\frac{\hat{\sigma}_{jk}}{\sqrt{n_{jk}}} \leq u_{ik} \leq \hat{s}_{jk} + C\frac{\hat{\sigma}_{jk}}{\sqrt{n_{jk}}}, \forall f_j \in a_i, \forall k. \quad (2)$$

where we look for t-table to find C corresponding to a certain confidence level(such as 95%) with the freedom of $n_{jk} - 1$. For simplicity, we represent the above confidence interval by $[\hat{s}_{jk} - d_{jk}, \hat{s}_{jk} + d_{jk}]$, where $d_{jk} = C\frac{\hat{\sigma}_{jk}}{\sqrt{n_{jk}}}$.

We introduce an indicator variable z_{ij} to represent whether the aspect phrase f_j belongs to aspect a_i, as follows:

$$z_{ji} = \begin{cases} 1 & ;\text{if } f_j \in a_i \\ 0 & ;\text{otherwise} \end{cases} \quad (3)$$

This leads to our SDC-constraint function.

$$\phi = z_{ji}|u_{ik} - \hat{s}_{jk}| \leq d_{jk}, \forall i, j, k \quad (4)$$

SDC-constraint are flexible for modeling Sentiment Distribution Consistency. The more observations we have, the smaller d_{jk} is. For frequent aspect phrase, the constraint can be very informative because it can filter unrelated aspects for aspect phrase f_j. The less observations we have, the larger d_{jk} is. For rare aspect phrases, the constraint can be very loose, and will not have much effect on the clustering process for aspect phrase f_j. In this way, the model can work very robustly.

SDC-constraints are data-driven constraints. Usually we have many reviews about hundreds of products in our dataset. For each aspect phrase, there are $|A| * |P|$ constraints (the number of aspects times the number of product). With thousands of constraints about which aspect it is not likely to belong to, the model learns to which aspect a phrase f_j should be assigned. Although most constraints may be loose because of the limited observations, SDC-constraint can still play an important role in the learning process.

2.3 Sentiment Distribution Consistency Regularized Multinomial Naive Bayes (SDC-MNB)

In this section, we present our probabilistic model which employs both context information and sentiment distribution.

First of all, we extract a context document d for each aspect phrase, which will be described in Section 2.5. In other word, a phrase is represented by its context document. Assuming that the documents in D are independent and identically distributed, the probability of generating D is then given by:

$$p_\theta(D) = \prod_{j=1}^{|D|} p_\theta(d_j) = \prod_{j=1}^{|D|} \sum_{y_j \in A} p_\theta(d_j, y_j) \quad (5)$$

where y_j is a latent variable indicating the aspect label for aspect phrase f_j, and θ is the model parameter.

In our problem, we are actually more interested in the posterior distribution over aspect, i.e., $p_\theta(y_j|d_j)$. Once the learned parameter θ is obtained, we can get our clustering result from $p_\theta(y_j|d_j)$, by assigning aspect a_i with the largest posterior to phrase f_j. We can also enforce SDC-constraint in expectation(on posterior p_θ). We use $q(Y)$ to denote the valid posterior distribution that satisfy our SDC-constraint, and Q to denote the valid posterior distribution space, as follows:

$$Q = \{q(Y) : E_q[z_{ji}|u_{ik} - \hat{s}_{jk}|] \leq d_{jk}, \forall i, j, k\}. \quad (6)$$

Since posterior plays such an important role in joining the context model and SDC-constraint, we formulate our problem in the framework of Posterior Regularization (PR). PR is an efficient framework to inject constraints on the posteriors of latent variables. Instead of restricting p_θ directly, which might not be feasible, PR penalizes the distance of p_θ to the constraint set Q. The posterior-regularized objective is termed as follows:

$$\max_\theta \{\log p_\theta(D) - \min_{q \in Q} KL(q(Y)||p_\theta(Y|D))\} \quad (7)$$

By trading off the data likelihood of the observed context documents (as defined in the first term), and the KL divergence of the posteriors to the valid posterior subspace defined by SDC-constraint (as defined in the second term), the objective encourages models with both desired posterior distribution and data likelihood. In essence, the model attempts to maximize data likelihood of context subject (softly) to SDC-constraint.

2.3.1 Multinomial Naive Bayes

In spirit to (Zhai et al., 2011a), we use Multinomial Naive Bayes (MNB) to model the context document. Let $w_{d_j,k}$ denotes the k^{th} word in document d_j, where each word is from the vocabulary $V = \{w_1, w_2, ..., w_{|V|}\}$. For each aspect phrase f_j, the probability of its latent aspect being a_i and generating context document d_i is

$$p_\theta(d_j, y_j = a_i) = p(a_i) \prod_{k=1}^{|d_j|} p(w_{d_j,k}|a_i) \quad (8)$$

where $p(a_i)$ and $p(w_{d_j,k}|a_i)$ are parameters of this model. Each word $w_{d_j,k}$ is conditionally independent of all other words given the aspect a_i.

Although MNB has been used in existing work for aspect clustering, all of the studies used it in a semi-supervised manner, with labeled data or pseudo-labeled data. In contrast, MNB proposed here is used in an unsupervised manner for aspect-related phrases clustering.

2.3.2 SDC-constraint

As mentioned above, the constraint posterior set Q is defined by

$$Q = \{q(Y) : q(y_j = a_i)|u_{ik} - \hat{s}_{jk}| \leq d_{jk}, \forall i, j, k\}. \quad (9)$$

We can see that Q denotes a set of linear constraints on the projected posterior distribution q. Note that we do not directly observe u_{ik}, the sentiment distribution of aspect a_i on product p_k. For aspect phrase f_j that belongs to aspect a_i, we estimate u_{ik} by counting all sentiment samples. We use the posterior $p_\theta(a_i|d_j)$ to approximately represent how likely phrase f_j belongs to aspect a_i.

$$u_{ik} = \frac{1}{\sum_{j=1}^{|D|} n_{jk} p_\theta(a_i|d_j)} \sum_{j=1}^{|D|} n_{jk} p_\theta(a_i|d_j) \hat{s}_{jk} \quad (10)$$

where $p_\theta(a_i|d_j)$ is short for $p_\theta(y_j = a_i|d_j)$, the probability that aspect phrase f_j belongs to a_i given the context document d_j. We estimate u_{ik} in this way because observations for aspect are relatively sufficient for a reliable estimation since observations for an aspect are aggregated from those for all phrases belonging to that aspect.

2.4 The Optimization Algorithm

The optimization algorithm for the objective (see Eq. 7) is an EM-like two-stage iterative algorithm.

In E-step, we first calculate the posterior distribution $p_\theta(a_i|d_j)$, then project it onto the valid posterior distribution space Q. Given the parameters θ, the posterior distribution can be calculated by Eq. 11.

$$p_\theta(a_i|d_j) = \frac{p(a_i) \prod_{k=1}^{|d_j|} p(w_{d_j,k}|a_i)}{\sum_{r=1}^{|A|} p(a_r) \prod_{k=1}^{|d_j|} p(w_{d_j,k}|a_r)} \quad (11)$$

We use the above posterior distribution to update the sentiment parameter for each aspect by Eq. 10. The projected posterior distribution q is calculated by

$$q = \underset{q \in Q}{argmin}\, \mathbf{KL}(q(Y)||p_\theta(Y|D)) \quad (12)$$

For each instance, there are $|A| * |P|$ constraints. However, we can prune a large number of useless constraints derived from limited observations. All constraints with $d_{jk} > 1$ can be pruned, due to the fact that the parameter u_{ik}, $\hat{s_{jk}}$ is within [0,1], and the difference can not be larger than 1. This optimization problem in Eq. 12 is easily solved via the dual form by the projected gradient algorithm (Boyd and Vandenberghe, 2004):

$$\max_{\lambda \geq 0} \Bigg(-\sum_{i=1}^{|A|} \sum_{k=1}^{|P|} \lambda_{ik} d_{jk} - \log \sum_{i=1}^{|A|} p_\theta(a_i|d_j) exp\{-\sum_{k=1}^{|P|} \lambda_{ik}|u_{ik} - \hat{s}_{jk}|\} - \epsilon\|\lambda\| \Bigg) \quad (13)$$

where ϵ controls the slack size for constraint. After solving the above optimization problem and obtaining the optimal λ, we can calculate the projected posterior distribution q by

$$q(y_j = a_i) = \frac{1}{Z} p_\theta(a_i|d_j) exp\{-\sum_{k=1}^{|P|} \lambda_{ik}|u_{ik} - \hat{s}_{jk}|\} \quad (14)$$

where Z is the normalization factor. Note that *sentiment distribution consistency* is actually modeled as instance-level constraint here, which makes it very efficient to solve.

In M-step, the projected posteriors $q(Y)$ are then used to compute sufficient statistics and update the models parameters θ. Given the projected posteriors $q(Y)$, the parameters can be updated by Eq. 15,16.

$$p(a_i) = \frac{1 + \sum_{j=1}^{|D|} q(y_j = a_i)}{|A| + |D|} \quad (15)$$

$$p(w_t|a_i) = \frac{1 + \sum_{j=1}^{|D|} N_{ti} q(y_j = a_i)}{|V| + \sum_{m=1}^{|V|} \sum_{j=1}^{|D|} N_{mj} q(y_j = a_i)} \quad (16)$$

where N_{tj} is the number of times that the word w_t occurs in document d_j.

The parameters are initialized randomly, and we repeat E-step and M-step until convergence.

2.5 Data Extraction

2.5.1 Context Extraction

In order to extract the context document d for each aspect phrase, we follow the approach in Zhai et al. (2011a). For each aspect phrase, we generate its context document by aggregating the surrounding texts of the phrase in all reviews. The preceding and following t words of a phrase are taken as the context where we set $t = 3$ in this paper. Stopwords and other aspect phrases are removed. For example, the following review contains two aspect phrases, "*screen*" and "*picture*",

The LCD ***screen*** *gives clear* ***picture***.

For "*screen*", the surrounding texts are {the, LCD, gives, clear, picture}. We remove stopwords "the", and the aspect term "*picture*", and the resultant context of "*screen*" in this review is

context(*screen*) ={LCD, screen, gives, clear}.

Similarly, the context of "*picture*" in this review is

context(*picture*) ={gives, clear}.

By aggregating the contexts of all the reviews that contain aspect phrase f_j, we obtain the corresponding context document d_j.

2.5.2 Sentiment Extraction

Since we use semi-structured reviews, we obtain the estimated sentiment distribution by simply counting how many times each aspect phrase appears in *Pros* and *Cons* reviews for each product respectively. So for each aspect phrase f_j, let n^+_{jk} denotes the times that f_j appears in *Pros* of all reviews for product p_k, and let n^-_{jk} denotes the times that f_j appears in *Cons* of all reviews for product p_k. So the total number of occurrence of a phrase is $n_{jk} = n^+_{jk} + n^-_{jk}$. We have samples like (1,1,1,0,0) where 1 means a phrase occurs in *Pros* of a review, and 0 in *Cons*. Given a sequence of such observations, the sample mean is easily computed as $\hat{s}_{jk} = \frac{n^+_{jk}}{n^+_{jk}+n^-_{jk}}$. And the sample standard deviation is $\hat{\sigma}_{jk} = \sqrt{\frac{(1-\hat{s}_{jk})^2 * n^+_{jk} + (\hat{s}_{jk})^2 * n^-_{jk}}{n_{jk}-1}}$.

3 Experiments

3.1 Data Preparation

The details of our review corpus are given in Table 2. This corpus contains semi-structured customer reviews from four domains: *Camera*, *Cellphone*, *Laptop*, and *MP3*. These reviews were crawled from the following web sites: www.amazon.cn, www.360buy.com, www.newegg.com.cn, and www.zol.com. The aspect label of each aspect phrases is annotated by human curators.

	Camera	Cellphone	Laptop	MP3
#Products	449	694	702	329
#Reviews	101,235	579,402	102,439	129,471
#Aspect Phrases	236	230	238	166
#Aspect	12	10	14	8

Table 2: Statistics of the review corpus. # denotes the size.

3.2 Evaluation Measures

We adapt three measures *Purity*, *Entropy*, and *Rand Index* for performance evaluation. These measures have been commonly used to evaluate clustering algorithms.

Given a data set DS, suppose its gold-standard partition is $G = \{g_1, ..., g_j, ..., g_k\}$, where k is the number of clusters. A clustering algorithm partitions DS into k disjoint subsets, say $DS_1, DS_2, ..., DS_k$.

Entropy: For each resulting cluster, we can measure its entropy using Eq. 17, where $P_i(g_j)$ is the proportion of data points of class g_j in DS_i. The entropy of the entire clustering result is calculated by Eq. 18.

$$entropy(DS_i) = -\sum_{j=1}^{k} P_i(g_j) log_2 P_i(g_j) \quad (17)$$

$$entropy(DS) = \sum_{i=1}^{k} \frac{|DS_i|}{|DS|} entropy(DS_i) \quad (18)$$

Purity: *Purity* measures the extent that a cluster contains only data from one gold-standard partition. The cluster purity is computed with Eq. 19. The total purity of the whole clustering result (all clusters) is computed with Eq. 20.

$$purity(DS_i) = \max_j P_i(g_j) \quad (19)$$

$$purity(DS) = \sum_{i=1}^{k} \frac{|DS_i|}{|DS|} purity(DS_i) \quad (20)$$

RI: The *Rand Index(RI)* penalizes both false positive and false negative decisions during clustering. Let TP (True Positive) denotes the number of pairs of elements that are in the same set in DS and in the same set in G. TN (True Negative) denotes number of pairs of elements that are in different sets in DS and in different sets in G. FP (False

	Camera			Cellphone			Laptop			MP3		
	P	RI	E	P	RI	E	P	RI	E	P	RI	E
Kmeans	43.48%	83.52%	2.098	48.91%	84.80%	1.792	43.46%	87.11%	2.211	40.00%	70.98%	2.047
L-EM	54.89%	87.07%	**1.690**	51.96%	86.64%	1.456	48.94%	84.53%	2.039	44.24%	75.37%	1.990
LDA	36.84%	83.28%	2.426	48.65%	85.33%	1.833	35.02%	83.53%	2.660	36.12%	76.08%	2.296
Constraint-LDA	43.30%	86.01%	2.216	47.89%	86.04%	1.974	32.35%	84.86%	2.676	50.70%	81.42%	1.924
SDC-MNB	**56.42%**	**88.16%**	1.725	**67.95%**	**90.62%**	**1.266**	**55.52%**	**90.72%**	**1.780**	**58.06%**	**83.57%**	**1.578**

Table 3: Comparison to unsupervised baselines. (P is short for purity, E for entropy, and RI for random index.)

Positive) denotes number of pairs of elements in S that are in the same set in DS and in different sets in G. FN (False Negative) denotes number of pairs of elements that are in different sets in DS and in the same set in G. The *Rand Index(RI)* is computed with Eq. 21.

$$RI(DS) = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

3.3 Evaluation Results

3.3.1 Comparison to unsupervised baselines

We compared our approach with several existing unsupervised methods. Some of the methods augmented unsupervised models by incorporating lexical similarity and other domain knowledge. All of them are context-based models.[2] We list these models as follows.

- ***Kmeans***: Kmeans is the most popular clustering algorithm. Here we use the context distributional similarity (cosine similarity) as the similarity measure.

- ***L-EM***: This is a state-of-the-art unsupervised method for clustering aspect phrases (Zhai et al., 2011a). L-EM employed lexical knowledge to provide a better initialization for EM.

- ***LDA***: LDA is a popular topic model(Blei et al., 2003). Given a set of documents, it outputs groups of terms of different topics. In our case, each aspect phrase is processed as a term. [3] Each sentence in a review is considered as a document. Each aspect is considered as a topic. In LDA, a term may belong to more than one topic/group, but we take the topic/group with the maximum probability.

- ***Constraint-LDA***: Constraint-LDA (Zhai et al., 2011b) is a state-of-the-art LDA-based method that incorporates must-link and cannot-link constraints for this task. We set the damping factor $\lambda = 0.3$ and relaxation factor $\eta = 0.9$, as suggested in the original reference.

For all methods that depend on the random initiation, we use the average results of 10 runs as the final result. For all LDA-based models, we choose $\alpha = 50/T$, $\beta = 0.1$, and run 1000 iterations.

Experiment results are shown in Table 3. We can see that our approach almost outperforms all unsupervised baseline methods by a large margin on all domains. In addition, we have the following observations:

- LDA and Kmeans perform poorly due to the fact that the two methods do not use any prior knowledge. It is also shown that only using the context distributional information is not sufficient for clustering aspect phrases.

- Constraint-LDA and L-EM that utilize prior knowledge perform better. We can see that Constraint-LDA outperforms LDA in terms of RI (Rand Index) on all domains. L-EM achieves the best results against the baselines. This demonstrates the effectiveness to incorporate prior knowledge.

- SDC-MNB produces the optimal results among all models for clustering. Methods that use must-links and cannot-links may suffer from noisy links. For L-EM, we find that it is sensitive to noisy must-links. As L-EM assumes that must-link is transitive, several noisy must-links may totally mislabel the softly annotated data. For Constraint-LDA, it is more robust than L-EM, because it doesn't assume the transitivity of must-link. However, it only promotes the RI (Rand Index) consistently by leveraging pair-wise prior knowledge, but sometimes it hurts the

[2] In our method, we collect context document for each aspect phrase. This process is conducted for L-EM and K-means. But for LDA and Constraint-LDA, we take each sentence of reviews as a document. This setting for the LDA baselines is adapted from previous work.

[3] Each aspect phrase is pre-processed as a single word (e.g., "battery life" is treated as battery-life). Other words are normally used in LDA.

performance with respect to purity or entropy. Our method is consistently better on almost all domains, which shows the advantages of the proposed model.

- SDC-MNB is remarkably better than baselines, particularly for the *cellphone* domain. We argue that this is because we have the largest number of reviews for each product in the *cellphone* domain. The larger dataset gives us more observations on each phrase, so that we obtain more reliable estimation of model parameters.

3.3.2 Comparison to supervised baselines

We further compare our methods with two supervised models. For each supervised model, we provide a proportion of manually labeled data for training, which is randomly selected from gold-standard annotations. However, we didn't use any labeled data for our approach.

- ***MNB***: The labeled seeds are used to train a MNB classifier to classify all unlabeled aspect phrases into different classes.

- ***L-Kmeans***: In L-Kmeans, the clusters of the labeled seeds are fixed at the initiation and remain unchanged during iteration.

	Purity	RI	Entropy
MNB-5%	53.21%	85.77%	1.854
MNB-10%	**59.55%**	**86.70%**	**1.656**
MNB-15%	66.06%	88.39%	1.449
L-Kmeans-10%	53.54%	86.15%	1.745
L-Kmeans-15%	**57.00%**	**86.89%**	**1.643**
L-Kmeans-20%	60.97%	87.63%	1.528
SDC-MNB	**59.49%**	**88.26%**	**1.580**

Table 4: Comparison to supervised baselines. MNB-5% means MNB with 5% labeled data.

We experiment with several settings: taking 5%, 10% and 15% of the manually labeled aspect phrases for training, and the remainder as unlabeled data. Experiment results is shown in Table 4 (the results are averaged over 4 domains). We can see that our unsupervised approach is roughly as good as the supervised MNB with 10% labeled data. Our unsupervised approach is also slightly better than L-Kmeans with 15% labeled data. This result further demonstrates the effectiveness of our model.

3.3.3 Influence of parameters

We vary the confidence level from 90% to 99.9% to see how it impacts on the performance of SDC-MNB. The results are presented in Fig. 5 (the results are averaged over 4 domains). We can see that the performance of clustering is fairly stable when changing the confidence level, which implies the robustness of our model.

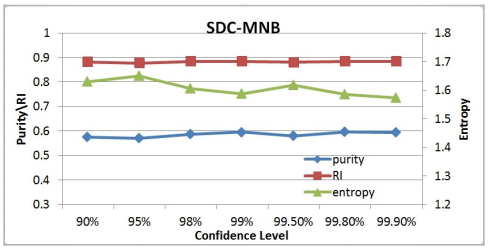


Figure 5: Influence of the confidence level on SDC-MNB.

3.3.4 Analysis of SDC-constraint

As mentioned in Section 2.2, SDC-constraint is dependent on the number of observations. More observations we get, more informative the constraint is, which means the constraint is tighter and d_{jk} (see Eq.4) is smaller. For all k, we count how many d_{jk} is less than 0.2 (and 1) on average for each aspect phrase f_j. d_{jk} is calculated with a confidence level of 99%. The statistics of constraints is given in Table 5. We can see that the cellphone domain has the most informative and largest constraint set, that may explain why SDC-MNB achieves the largest purity gain(over L-EM) in cellphone domain.

	$\#(d_{jk} < 0.2)$	$\#(0.2 < d_{jk} < 1)$	purity gain
Camera	3.02	8.78	1.53%
Cellphone	17.29	30.5	15.99%
Laptop	4.6	13.22	6.58%
MP3MP4	6.1	10.7	13.82%

Table 5: Constraint statistics on different domains.

4 Related Work

Our work is related to two important research topics: aspect-level sentiment analysis, and constraint-driven learning. For aspect-level sentiment analysis, aspect extraction and clustering are key tasks. For constraint-driven learning, a variety of frameworks and models for sentiment analysis have been studied extensively.

There have been many studies on clustering aspect-related phrases. Most existing studies are

based on context information. Some works also encoded lexical similarity and synonyms as prior knowledge. Carenini et al. (2005) proposed a method that was based on several similarity metrics involving string similarity, synonyms, and lexical distances defined with WordNet. Guo et al. (2009) proposed a multi-level latent semantic association model to capture expression-level and context-level topic structure. Zhai et al. (2010) proposed an EM-based semi-supervised learning method to group aspect expressions into user-specified aspects. They employed lexical knowledge to provide a better initialization for EM. In Zhai et al. (2011a), an EM-based unsupervised version was proposed. The so-called L-EM model first generated softly labeled data by grouping feature expressions that share words in common, and then merged the groups by lexical similarity. Zhai et al. (2011b) proposed a LDA-based method that incorporates must-link and cannot-link constraints.

Another line of work aimed to extract and cluster aspect words simultaneously using topic modeling. Titov and McDonald (2008) proposed the multi-grain topic models to discover global and local aspects. Branavan et al. (2008) proposed a method which first clustered the key-phrases in Pros and Cons into some aspect categories based on distributional similarity, then built a topic model modeling the topics or aspects. Zhao et al. (2010) proposed the MaxEnt-LDA (a Maximum Entropy and LDA combination) hybrid model to jointly discover both aspect words and aspect-specific opinion words, which can leverage syntactic features to separate aspects and sentiment words. Mukherjee and Liu (2012) proposed a semi-supervised topic model which used user-provided seeds to discover aspects. Chen et al. (2013) proposed a knowledge-based topic model to incorporate must-link and cannot-link information. Their model can adjust topic numbers automatically by leveraging cannot-link.

Our work is also related to general constraint-driven(or knowledge-driven) learning models. Several general frameworks have been proposed to fully utilize various prior knowledge in learning. Constraint-driven learning (Chang et al., 2008) (CODL) is an EM-like algorithm that incorporates per-instance constraints into semi-supervised learning. Posterior regularization (Graca et al., 2007) (PR) is a modified EM algorithm in which the E-step is replaced by the projection of the model posterior distribution onto the set of distributions that satisfy auxiliary expectation constraints. Generalized expectation criteria (Druck et al., 2008) (GE) is a framework for incorporating preferences about model expectations into parameter estimation objective functions. Liang et al. (2009) developed a Bayesian decision-theoretic framework to learn an exponential family model using general measurements on the unlabeled data. In this paper, we model our problem in the framework of posterior regularization.

Many works promoted the performance of sentiment analysis by incorporating prior knowledge as weak supervision. Li and Zhang (2009) injected lexical prior knowledge to non-negative matrix tri-factorization. Shen and Li (2011) further extended the matrix factorization framework to model dual supervision from both document and word labels. Vikas Sindhwani (2008) proposed a general framework for incorporating lexical information as well as unlabeled data within standard regularized least squares for sentiment prediction tasks. Fang (2013)proposed a structural learning model with a handful set of aspect signature terms that are encoded as weak supervision to extract latent sentiment explanations.

5 Conclusions

Aspect finding and clustering is an important task for aspect-level sentiment analysis. In order to cluster aspect-related phrases, this paper has explored a novel concept, *sentiment distribution consistency*. We formalize the concept as soft constraint, integrate the constraint with a context-based probabilistic model, and solve the problem in the posterior regularization framework. The proposed model is also designed to be robust with both sufficient and insufficient observations. Experiments show that our approach outperforms state-of-the-art baselines consistently.

Acknowledgments

This work was partly supported by the following grants from: the National Basic Research Program (973 Program) under grant No.2012CB316301 and 2013CB329403, the National Science Foundation of China project under grant No.61332007 and No. 61272227, and the Beijing Higher Education Young Elite Teacher Project.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, New York, NY, USA.

S. R. K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Giuseppe Carenini, Raymond T. Ng, and Ed Zwart. 2005. Extracting knowledge from evaluative text. In *Proceedings of the 3rd International Conference on Knowledge Capture*, K-CAP '05, pages 11–18, New York, NY, USA. ACM.

Ming-Wei Chang, Lev Ratinov, Nicholas Rizzolo, and Dan Roth. 2008. Learning and inference with constraints. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1513–1518. AAAI Press.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Mal Castellanos, and Riddhiman Ghosh. 2013. Exploiting domain knowledge in aspect extraction. In *EMNLP*, pages 1655–1667. ACL.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 595–602, New York, NY, USA. ACM.

Lei Fang, Minlie Huang, and Xiaoyan Zhu. 2013. Exploring weakly supervised latent sentiment explanations for aspect-level review analysis. In Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi, editors, *CIKM*, pages 1057–1066. ACM.

Joao V. Graca, Lf Inesc-id, Kuzman Ganchev, Ben Taskar, Joo V. Graa, L F Inesc-id, Kuzman Ganchev, and Ben Taskar. 2007. Expectation maximization and posterior constraints. In *In Advances in NIPS*, pages 569–576.

Honglei Guo, Huijia Zhu, Zhili Guo, XiaoXun Zhang, and Zhong Su. 2009. Product feature categorization with multilevel latent semantic association. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1087–1096, New York, NY, USA. ACM.

Tao Li, Yi Zhang, and Vikas Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 244–252, Stroudsburg, PA, USA. Association for Computational Linguistics.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 641–648, New York, NY, USA. ACM.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 339–348, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chao Shen and Tao Li. 2011. A non-negative matrix factorization based approach for active dual supervision from document and word labels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 949–958, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vikas Sindhwani and Prem Melville. 2008. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*, pages 1025–1030. IEEE Computer Society.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA. ACM.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2010. Grouping product features using semi-supervised learning with soft-constraints. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1272–1280, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011a. Clustering product features for opinion mining. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 347–354, New York, NY, USA. ACM.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011b. Constrained lda for grouping product features in opinion mining. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, PAKDD'11, pages 448–459, Berlin, Heidelberg. Springer-Verlag.

Wayne X. Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65, Stroudsburg, PA, USA. Association for Computational Linguistics.