Aspect Extraction with Automated Prior Knowledge Learning

Zhiyuan Chen Arjun Mukherjee Bing Liu
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
{czyuanacm,arjun4787}@gmail.com,liub@cs.uic.edu

Abstract

Aspect extraction is an important task in sentiment analysis. Topic modeling is a popular method for the task. However, unsupervised topic models often generate incoherent aspects. To address the issue, several knowledge-based models have been proposed to incorporate prior knowledge provided by the user to guide modeling. In this paper, we take a major step forward and show that in the big data era, without any user input, it is possible to learn prior knowledge automatically from a large amount of review data available on the Web. Such knowledge can then be used by a topic model to discover more coherent aspects. There are two key challenges: (1) learning quality knowledge from reviews of diverse domains, and (2) making the model fault-tolerant to handle possibly wrong knowledge. A novel approach is proposed to solve these problems. Experimental results using reviews from 36 domains show that the proposed approach achieves significant improvements over state-of-the-art baselines.

1 Introduction

Aspect extraction aims to extract target entities and their aspects (or attributes) that people have expressed opinions upon (Hu and Liu, 2004, Liu, 2012). For example, in "*The voice is not clear*," the aspect term is "*voice*." Aspect extraction has two subtasks: *aspect term extraction* and *aspect term resolution*. Aspect term resolution groups extracted synonymous aspect terms together. For example, "voice" and "sound" should be grouped together as they refer to the same aspect of phones.

Recently, topic models have been extensively applied to aspect extraction because they can perform both subtasks at the same time while other existing methods all need two separate steps (see Section 2). Traditional topic models such as LDA (Blei et al., 2003) and pLSA (Hofmann, 1999) are unsupervised methods for extracting latent topics in text documents. Topics are aspects in our task. Each aspect (or topic) is a distribution over (aspect) terms. However, researchers have shown that fully unsupervised models often produce incoherent topics because the objective functions of topic models do not always correlate well with human judgments (Chang et al., 2009).

To tackle the problem, several semi-supervised topic models, also called knowledge-based topic models, have been proposed. DF-LDA (Andrzejewski et al., 2009) can incorporate two forms of prior knowledge from the user: *must-links* and *cannot-links*. A must-link implies that two terms (or words) should belong to the same topic whereas a cannot-link indicates that two terms should not be in the same topic. In a similar but more generic vein, *must-sets* and *cannot-sets* are used in MC-LDA (Chen et al., 2013b). Other related works include (Andrzejewski et al., 2011, Chen et al., 2013a, Chen et al., 2013c, Mukherjee and Liu, 2012, Hu et al., 2011, Jagarlamudi et al., 2012, Lu et al., 2011, Petterson et al., 2010). They all allow prior knowledge to be specified by the user to guide the modeling process.

In this paper, we take a major step further. We mine the prior knowledge directly from a large amount of relevant data without any user intervention, and thus make this approach fully automatic. We hypothesize that it is possible to learn quality prior knowledge from the big data (of reviews) available on the Web. The intuition is that although every domain is different, there is a decent amount of aspect overlapping across domains. For example, every product domain has the aspect/topic of "*price*," most electronic products share the aspect "*battery*" and some also share "*screen*." Thus, the shared aspect knowl-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 347–358,
Baltimore, Maryland, USA, June 23-25 2014.

edge mined from a set of domains can potentially help improve aspect extraction in each of these domains, as well as in new domains. Our proposed method aims to achieve this objective. There are two major challenges: (1) learning quality knowledge from a large number of domains, and (2) making the extraction model fault-tolerant, i.e., capable of handling possibly incorrect learned knowledge. We briefly introduce the proposed method below, which consists of two steps.

Learning quality knowledge: Clearly, learned knowledge from only a single domain can be erroneous. However, if the learned knowledge is shared by multiple domains, the knowledge is more likely to be of high quality. We thus propose to first use LDA to learn topics/aspects from each individual domain and then discover the shared aspects (or topics) and aspect terms among a subset of domains. These shared aspects and aspect terms are more likely to be of good quality. They can serve as the prior knowledge to guide a model to extract aspects. A piece of knowledge is a set of semantically coherent (aspect) terms which are likely to belong to the same topic or aspect, i.e., similar to a must-link, but mined automatically.

Extraction guided by learned knowledge: For reliable aspect extraction using the learned prior knowledge, we must account for possible errors in the knowledge. In particular, a piece of automatically learned knowledge may be wrong or domain specific (i.e., the words in the knowledge are semantically coherent in some domains but not in others). To leverage such knowledge, the system must detect those inappropriate pieces of knowledge. We propose a method to solve this problem, which also results in a new topic model, called AKL (*Automated Knowledge LDA*), whose inference can exploit the automatically learned prior knowledge and handle the issues of incorrect knowledge to produce superior aspects.

In summary, this paper makes the following contributions:

1. It proposes to exploit the big data to learn prior knowledge and leverage the knowledge in topic models to extract more coherent aspects. The process is fully automatic. To the best of our knowledge, none of the existing models for aspect extraction is able to achieve this.
2. It proposes an effective method to learn quality knowledge from raw topics produced using review corpora from many different domains.
3. It proposes a new inference mechanism for topic modeling, which can handle incorrect knowledge in aspect extraction.

2 Related Work

Aspect extraction has been studied by many researchers in sentiment analysis (Liu, 2012, Pang and Lee, 2008), e.g., using supervised sequence labeling or classification (Choi and Cardie, 2010, Jakob and Gurevych, 2010, Kobayashi et al., 2007, Li et al., 2010, Yang and Cardie, 2013) and using word frequency and syntactic patterns (Hu and Liu, 2004, Ku et al., 2006, Liu et al., 2013, Popescu and Etzioni, 2005, Qiu et al., 2011, Somasundaran and Wiebe, 2009, Wu et al., 2009, Xu et al., 2013, Yu et al., 2011, Zhao et al., 2012, Zhou et al., 2013, Zhuang et al., 2006). However, these works only perform extraction but not aspect term grouping or resolution. Separate aspect term grouping has been done in (Carenini et al., 2005, Guo et al., 2009, Zhai et al., 2011). They assume that aspect terms have been extracted beforehand.

To extract and group aspects simultaneously, topic models have been applied by researchers (Branavan et al., 2008, Brody and Elhadad, 2010, Chen et al., 2013b, Fang and Huang, 2012, He et al., 2011, Jo and Oh, 2011, Kim et al., 2013, Lazaridou et al., 2013, Li et al., 2011, Lin and He, 2009, Lu et al., 2009, Lu et al., 2012, Lu and Zhai, 2008, Mei et al., 2007, Moghaddam and Ester, 2013, Mukherjee and Liu, 2012, Sauper and Barzilay, 2013, Titov and McDonald, 2008, Wang et al., 2010, Zhao et al., 2010). Our proposed AKL model belongs to the class of knowledge-based topic models. Besides the knowledge-based topic models discussed in Section 1, document labels are incorporated as implicit knowledge in (Blei and McAuliffe, 2007, Ramage et al., 2009). Geographical region knowledge has also been considered in topic models (Eisenstein et al., 2010). All of these models assume that the prior knowledge is correct. GK-LDA (Chen et al., 2013a) is the only knowledge-based topic model that deals with wrong lexical knowledge to some extent. As we will see in Section 6, AKL outperformed GK-LDA significantly due to AKL's more effective error handling mechanism. Furthermore, GK-LDA does not learn any prior knowledge.

Our work is also related to transfer learning to some extent. Topic models have been used to help

Input: Corpora $\boldsymbol{D_L}$ for knowledge learning
Test corpora $\boldsymbol{D_T}$

1: // STEP 1: Learning prior knowledge.
2: **for** $r = 0$ **to** R **do** // Iterate $R + 1$ times.
3: **for** each domain corpus $D_i \in \boldsymbol{D_L}$ **do**
4: **if** $r = 0$ **then**
5: $A_i \leftarrow \text{LDA}(D_i)$;
6: **else**
7: $A_i \leftarrow \text{AKL}(D_i, \boldsymbol{K})$;
8: **end if**
9: **end for**
10: $\boldsymbol{A} \leftarrow \cup_i A_i$;
11: $TC \leftarrow \text{Clustering}(\boldsymbol{A})$;
12: **for** each cluster $T_j \in TC$ **do**
13: $K_j \leftarrow \text{FPM}(T_j)$;
14: **end for**
15: $\boldsymbol{K} \leftarrow \cup_j K_j$;
16: **end for**
17: // STEP 2: Using the learned knowledge.
18: **for** each test corpus $D_i \in \boldsymbol{D_T}$ **do**
19: $A_i \leftarrow \text{AKL}(D_i, \boldsymbol{K})$;
20: **end for**

Figure 1: The proposed overall algorithm.

transfer learning (He et al., 2011, Pan and Yang, 2010, Xue et al., 2008). However, transfer learning in these papers is for traditional classification rather than topic/aspect extraction. In (Kang et al., 2012), labeled documents from source domains are transferred to the target domain to produce topic models with better fitting. However, we do not use any labeled data. In (Yang et al., 2011), a user provided parameter indicating the technicality degree of a domain was used to model the language gap between topics. In contrast, our method is fully automatic without human intervention.

3 Overall Algorithm

This section introduces the proposed overall algorithm. It consists of two main steps: *learning quality knowledge* and *using the learned knowledge*. Figure 1 gives the algorithm.

Step 1 (learning quality knowledge, Lines 1-16): The input is the review corpora $\boldsymbol{D_L}$ from multiple domains, from which the knowledge is automatically learned. Lines 3 and 5 run LDA on each review domain corpus $D_i \in \boldsymbol{D_L}$ to generate a set of aspects/topics A_i (lines 2, 4, and 6-9 will be discussed below). Line 10 unions the topics from all domains to give $\boldsymbol{A}$. Lines 11-14 cluster the topics in $\boldsymbol{A}$ into some coherent groups (or clusters) and then discover knowledge K_j from each group of topics using frequent pattern mining (FPM) (Han et al., 2007). We will detail these in Section 4. Each piece of the learned knowledge is a set of terms which are likely to belong to the same aspect.

Iterative improvement: The above process can actually run iteratively because the learned knowledge $\boldsymbol{K}$ can help the topic model learn better topics in each domain $D_i \in \boldsymbol{D_L}$, which results in better knowledge $\boldsymbol{K}$ in the next iteration. This iterative process is reflected in lines 2, 4, 6-9 and 16. We will examine the performance of the process at different iterations in Section 6.2. From the second iteration, we can use the knowledge learned from the previous iteration (lines 6-8). The learned knowledge is leveraged by the new model AKL, which is discussed below in Step 2.

Step 2 (using the learned knowledge, Lines 17-20): The proposed model AKL is employed to use the learned knowledge $\boldsymbol{K}$ to help topic modeling in test domains $\boldsymbol{D_T}$, which can be $\boldsymbol{D_L}$ or other unseen domains. The key challenge of this step is how to use the learned prior knowledge $\boldsymbol{K}$ effectively in AKL and deal with possible errors in $\boldsymbol{K}$. We will elaborate them in Section 5.

Scalability: the proposed algorithm is naturally scalable as both LDA and AKL run on each domain independently. Thus, for all domains, the algorithm can run in parallel. Only the resulting topics need to be brought together for knowledge learning (Step 1). These resulting topics used in learning are much smaller than the domain corpus as only a list of top terms from each topic are utilized due to their high reliability.

4 Learning Quality Knowledge

This section details Step 1 in the overall algorithm, which has three sub-steps: running LDA (or AKL) on each domain corpus, clustering the resulting topics, and mining frequent patterns from the topics in each cluster. Since running LDA is simple, we will not discuss it further. The proposed AKL model will be discussed in Section 5. Below we focus on the other two sub-steps.

4.1 Topic Clustering

After running LDA (or AKL) on each domain corpus, a set of topics is obtained. Each topic is a distribution over terms (or words), i.e., terms with their associated probabilities. Here, we use only the top terms with high probabilities. As discussed earlier, quality knowledge should be shared

by topics across several domains. Thus, it is natural to exploit a frequency-based approach to discover frequent set of terms as quality knowledge. However, we need to deal with two issues.

1. Generic aspects, such as *price* with aspect terms like *cost* and *pricy*, are shared by many (even all) product domains. But specific aspects such as *screen*, occur only in domains with products having them. It means that different aspects may have distinct frequencies. Thus, using a single frequency threshold in the frequency-based approach is not sufficient to extract both generic and specific aspects because the generic aspects will result in numerous spurious aspects (Han et al., 2007).
2. A term may have multiple senses in different domains. For example, *light* can mean "of little weight" or "something that makes things visible". A good knowledge base should have the capacity of handling this ambiguity.

To deal with these two issues, we propose to discover knowledge in two stages: topic clustering and frequent pattern mining (FPM).

The purpose of clustering is to group raw topics from a topic model (LDA or AKL) into clusters. Each cluster contains semantically related topics likely to indicate the same real-world aspect. We then mine knowledge from each cluster using an FPM technique. Note that the multiple senses of a term can be distinguished by the semantic meanings represented by the topics in different clusters.

For clustering, we tried k-means and k-medoids (Kaufman and Rousseeuw, 1990), and found that k-medoids performs slightly better. One possible reason is that k-means is more sensitive to outliers. In our topic clustering, each data point is a topic represented by its top terms (with their probabilities normalized). The distance between two data points is measured by symmetrised KL-Divergence.

4.2 Frequent Pattern Mining

Given topics within each cluster, this step finds sets of terms that appear together in multiple topics, i.e., shared terms among similar topics across multiple domains. Terms in such a set are likely to belong to the same aspect. To find such sets of terms within each cluster, we use *frequent pattern mining* (FPM) (Han et al., 2007), which is suited for the task. The probability of each term is ignored in FPM.

FPM is stated as follows: Given a set of transactions $\mathbb{T}$, where each transaction $t_i \in \mathbb{T}$ is a set of items from a global item set $\mathbb{I}$, i.e., $t_i \in \mathbb{I}$. In our context, t_i is the topic vector comprising the top terms of a topic (no probability attached). $\mathbb{T}$ is the collection of all topics within a cluster and $\mathbb{I}$ is the set of all terms in $\mathbb{T}$. The goal of FPM is to find all patterns that satisfy some user-specified frequency threshold (also called minimum support count), which is the minimum number of times that a pattern should appear in $\mathbb{T}$. Such patterns are called frequent patterns. In our context, a pattern is a set of terms which have appeared multiple times in the topics within a cluster. Such patterns compose our knowledge base as shown below.

4.3 Knowledge Representation

As the knowledge is extracted from each cluster individually, we represent our knowledge base as a set of clusters, where each cluster consists of a set of frequent 2-patterns mined using FPM, e.g.,

Cluster 1: {battery, life}, {battery, hour}, {battery, long}, {charge, long}

Cluster 2: {service, support}, {support, customer}, {service, customer}

Using two terms in a set is sufficient to cover the semantic relationship of the terms belonging to the same aspect. Longer patterns tend to contain more errors since some terms in a set may not belong to the same aspect as others. Such partial errors hurt performance in the downstream model.

5 AKL: Using the Learned Knowledge

We now present the proposed topic model AKL, which is able to use the automatically learned knowledge to improve aspect extraction.

5.1 Plate Notation

Differing from most topic models based on topic-term distribution, AKL incorporates a latent cluster variable $\boldsymbol{c}$ to connect topics and terms. The plate notation of AKL is shown in Figure 2. The inputs of the model are M documents, T topics and C clusters. Each document m has N_m terms. We model distribution $P(cluster|topic)$ as ψ and distribution $P(term|topic, cluster)$ as φ with Dirichlet priors β and γ respectively. $P(topic|document)$ is modeled by θ with a Dirichlet prior α. The terms in each document are assumed to be generated by first sampling a topic z, and then a cluster c given topic z, and finally

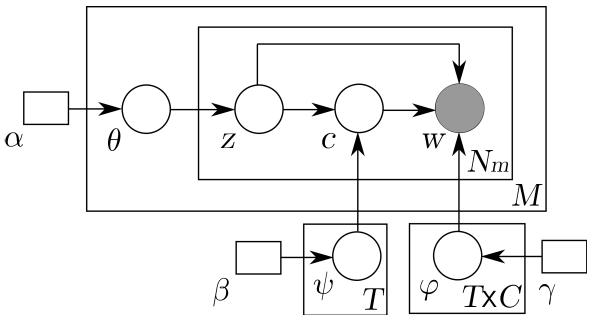

Figure 2: Plate notation for AKL.

a term w given topic z and cluster c. This plate notation of AKL and its associated generative process are similar to those of MC-LDA (Chen et al., 2013b). However, there are three key differences.

1. Our knowledge is automatically mined which may have errors (or noises), while the prior knowledge for MC-LDA is manually provided and assumed to be correct. As we will see in Section 6, using our knowledge, MC-LDA does not generate as coherent aspects as AKL.
2. Our knowledge is represented as clusters. Each cluster contains a set of frequent 2-patterns with semantically correlated terms. They are different from must-sets used in MC-LDA.
3. Most importantly, due to the use of the new form of knowledge, AKL's inference mechanism (Gibbs sampler) is entirely different from that of MC-LDA (Section 5.2), which results in superior performances (Section 6). Note that the inference mechanism and the prior knowledge cannot be reflected in the plate notation for AKL in Figure 2.

In short, our modeling contributions are (1) the capability of handling more expressive knowledge in the form of clusters, (2) a novel Gibbs sampler to deal with inappropriate knowledge.

5.2 The Gibbs Sampler

As the automatically learned prior knowledge may contain errors for a domain, AKL has to learn the usefulness of each piece of knowledge dynamically during inference. Instead of assigning weights to each piece of knowledge as a fixed prior in (Chen et al., 2013a), we propose a new Gibbs sampler, which can dynamically balance the use of prior knowledge and the information in the corpus during the Gibbs sampling iterations.

We adopt a Blocked Gibbs sampler (Rosen-Zvi et al., 2010) as it improves convergence and reduces autocorrelation when the variables (topic z and cluster c in AKL) are highly related. For each term w_i in each document, we jointly sample a topic z_i and cluster c_i (containing w_i) based on the conditional distribution in Gibbs sampler (will be detailed in Equation 4). To compute this distribution, instead of considering how well z_i matches with w_i only (as in LDA), we also consider two other factors:

1. The extent c_i corroborates w_i given the corpus. By "corroborate", we mean whether those frequent 2-patterns in c_i containing w_i are also supported by the actual information in the domain corpus to some extent (see the measure in Equation 1 below). If c_i corroborates w_i well, c_i is likely to be useful, and thus should also provide guidance in determining z_i. Otherwise, c_i may not be a suitable piece of knowledge for w_i in the domain.
2. Agreement between c_i and z_i. By agreement we mean the degree that the terms (union of all frequent 2-patterns of c_i) in cluster c_i are reflected in topic z_i. Unlike the first factor, this is a global factor as it concerns all the terms in a knowledge cluster.

For the first factor, we measure how well c_i corroborates w_i given the corpus based on co-document frequency ratio. As shown in (Mimno et al., 2011), co-document frequency is a good indicator of term correlation in a domain. Following (Mimno et al., 2011), we define a symmetric co-document frequency ratio as follows:

$$Co\text{-}Doc(w, w') = \frac{D(w, w') + 1}{(D(w) + D(w')) \times \frac{1}{2} + 1} \quad (1)$$

where (w, w') refers to each frequent 2-pattern in the knowledge cluster c_i. $D(w, w')$ is the number of documents that contain both terms w and w' and $D(w)$ is the number of documents containing w. A smoothing count of 1 is added to avoid the ratio being 0.

For the second factor, if cluster c_i and topic z_i agree, the intuition is that the terms in c_i (union of all frequent 2-patterns of c_i) should appear as top terms under z_i (i.e., ranked top according to the term probability under z_i). We define the agreement using symmetrised KL-Divergence between the two distributions ($DIST_c$ and $DIST_z$) corresponding to c_i and z_i respectively. As there is no prior preference on the terms of c_i, we use the uniform distribution over all terms in c_i for $DIST_c$. For $DIST_z$, as only top 20 terms under z_i are usually reliable, we use these top terms

with their probabilities (re-normalized) to represent the topic. Note that a smoothing probability (i.e., a very small value) is also given to every term for calculating KL-Divergence. Given $DIST_c$ and $DIST_z$, the agreement is computed with:

$$Agreement(c, z) = \frac{1}{KL(DIST_c, DIST_z)} \quad (2)$$

The rationale of Equation 2 is that the lesser divergence between $DIST_c$ and $DIST_z$ implies the more agreement between c_i and z_i.

We further employ the *Generalized Plya urn* (GPU) model (Mahmoud, 2008) which was shown to be effective in leveraging semantically related words (Chen et al., 2013a, Chen et al., 2013b, Mimno et al., 2011). The GPU model here basically states that assigning topic z_i and cluster c_i to term w_i will not only increase the probability of connecting z_i and c_i with w_i, but also make it more likely to associate z_i and c_i with term w' where w' shares a 2-pattern with w_i in c_i. The amount of probability increase is determined by matrix $\mathbb{A}_{c,w',w}$ defined as:

$$\mathbb{A}_{c,w',w} = \begin{cases} 1, & \text{if } w = w' \\ \sigma, & \text{if } (w, w') \in c, w \neq w' \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where value 1 controls the probability increase of w by seeing w itself, and σ controls the probability increase of w' by seeing w. Please refer to (Chen et al., 2013b) for more details.

Putting together Equations 1, 2 and 3 into a blocked Gibbs Sampler, we can define the following sampling distribution in Gibbs sampler so that it provides helpful guidance in determining the usefulness of the prior knowledge and in selecting the semantically coherent topic.

$$\begin{aligned} &P(z_i = t, c_i = c | \boldsymbol{z}^{-i}, \boldsymbol{c}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma, \mathbb{A}) \\ &\propto \sum_{(w,w') \in c} Co\text{-}Doc(w, w') \times Agreement(c, t) \\ &\times \frac{n^{-i}_{m,t} + \alpha}{\sum_{t'=1}^{T}(n^{-i}_{m,t'} + \alpha)} \\ &\times \frac{\sum_{w'=1}^{V}\sum_{v'=1}^{V} \mathbb{A}_{c,v',w'} \times n^{-i}_{t,c,v'} + \beta}{\sum_{c'=1}^{C}(\sum_{w'=1}^{V}\sum_{v'=1}^{V} \mathbb{A}_{c',v',w'} \times n^{-i}_{t,c',v'} + \beta)} \\ &\times \frac{\sum_{w'=1}^{V} \mathbb{A}_{c,w',w_i} \times n^{-i}_{t,c,w'} + \gamma}{\sum_{v'=1}^{V}(\sum_{w'=1}^{V} \mathbb{A}_{c,w',v'} \times n^{-i}_{t,c,w'} + \gamma)} \end{aligned} \quad (4)$$

where n^{-i} denotes the count excluding the current assignment of z_i and c_i, i.e., $\boldsymbol{z}^{-i}$ and $\boldsymbol{c}^{-i}$. $n_{m,t}$ denotes the number of times that topic t was assigned to terms in document m. $n_{t,c}$ denotes the times that cluster c occurs under topic t. $n_{t,c,v}$ refers to the number of times that term v appears in cluster c under topic t. α, β and γ are predefined Dirichlet hyperparameters.

Note that although the above Gibbs sampler is able to distinguish useful knowledge from wrong knowledge, it is possible that there is no cluster corroborates for a particular term. For every term w, apart from its knowledge clusters, we also add a singleton cluster for w, i.e., a cluster with one pattern $\{w, w\}$ only. When no knowledge cluster is applicable, this singleton cluster is used. As a singleton cluster does not contain any knowledge information but only the word itself, Equations 1 and 2 cannot be computed. For the values of singleton clusters for these two equations, we assign them as the averages of those values of all non-singleton knowledge clusters.

6 Experiments

This section evaluates and compares the proposed AKL model with three baseline models LDA, MC-LDA, and GK-LDA. LDA (Blei et al., 2003) is the most popular unsupervised topic model. MC-LDA (Chen et al., 2013b) is a recent knowledge-based model for aspect extraction. GK-LDA (Chen et al., 2013a) handles wrong knowledge by setting prior weights using the ratio of word probabilities. Our automatically extracted knowledge is provided to these models. Note that cannot-set of MC-LDA is not used in AKL.

6.1 Experimental Settings

Dataset. We created a large dataset containing reviews from 36 product domains or types from Amazon.com. The product domain names are listed in Table 1. Each domain contains $1,000$ reviews. This gives us 36 domain corpora. We have made the dataset publically available at the website of the first author.

Pre-processing. We followed (Chen et al., 2013b) to employ standard pre-processing like lemmatization and stopword removal. To have a fair comparison, we also treat each sentence as a document as in (Chen et al., 2013a, Chen et al., 2013b).

Parameter Settings. For all models, posterior estimates of latent variables were taken with a sampling lag of 20 iterations in the post burn-in phase (first 200 iterations for burn-in) with $2,000$ iterations in total. The model parameters were tuned on the development set in our pilot experiments

Amplifier	DVD Player	Kindle	MP3 Player	Scanner	Video Player
Blu-Ray Player	GPS	Laptop	Network Adapter	Speaker	Video Recorder
Camera	Hard Drive	Media Player	Printer	Subwoofer	Watch
CD Player	Headphone	Microphone	Projector	Tablet	Webcam
Cell Phone	Home Theater System	Monitor	Radar Detector	Telephone	Wireless Router
Computer	Keyboard	Mouse	Remote Control	TV	Xbox

Table 1: List of 36 domain names.

and set to $\alpha = 1$, $\beta = 0.1$, $T = 15$, and $\sigma = 0.2$. Furthermore, for each cluster, γ is set proportional to the number of terms in it. The other parameters for MC-LDA and GK-LDA were set as in their original papers. For parameters of AKL, we used the top 15 terms for each topic in the clustering phrase. The number of clusters is set to the number of domains. We will test the sensitivity of these clustering parameters in Section 6.4. The minimum support count for frequent pattern mining was set empirically to $min(5, 0.4 \times \#\mathbb{T})$, where $\#\mathbb{T}$ is the number of transactions (i.e., the number of topics from all domains) in a cluster.

Test Settings: We use two test settings as below:

1. (Sections 6.2, 6.3 and 6.4) Test on the same corpora as those used in learning the prior knowledge. This is meaningful as the learning phrase is automatic and unsupervised (Figure 1). We call this *self-learning-and-improvement*.
2. (Section 6.5) Test on new/unseen domain corpora after knowledge learning.

6.2 Topic Coherence

This sub-section evaluates the topics/aspects generated by each model based on Topic Coherence (Mimno et al., 2011) in test setting 1. Traditionally, topic models have been evaluated using perplexity. However, perplexity on the held-out test set does not reflect the semantic coherence of topics and may be contrary to human judgments (Chang et al., 2009). Instead, the metric Topic Coherence has been shown in (Mimno et al., 2011) to correlate well with human judgments. Recently, it has become a standard practice to use Topic Coherence for evaluation of topic models (Arora et al., 2013). A higher Topic Coherence value indicates a better topic interpretability, i.e., semantically more coherent topics.

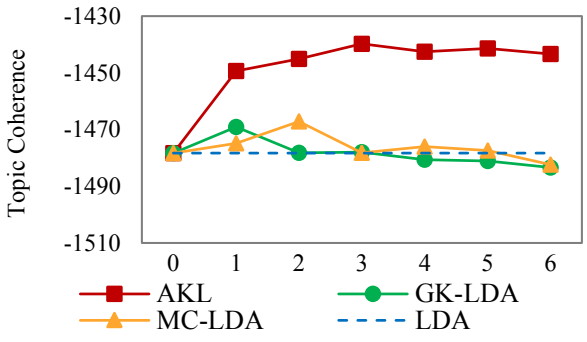

Figure 3: Average Topic Coherence of each model at different learning iterations (Iteration 0 is equivalent to LDA).

Figure 3 shows the average Topic Coherence of each model using knowledge learned at different learning iterations (Figure 1). For MC-LDA or GK-LDA, this is done by replacing AKL in lines 7 and 19 of Figure 1 with MC-LDA or GK-LDA. Each value is the average over all 36 domains. From Figure 3, we can observe the followings:

1. AKL performs the best with the highest Topic Coherence values at all iterations. It is actually the best in all 36 domains. These show that AKL finds more interpretable topics than the baselines. Its values stabilize after iteration 3.
2. Both GK-LDA and MC-LDA perform slightly better than LDA in iterations 1 and 2. MC-LDA does not handle wrong knowledge. This shows that the mined knowledge is of good quality. Although GK-LDA uses large word probability differences under a topic to detect wrong lexical knowledge, it is not as effective as AKL. The reason is that as the lexical knowledge is from general dictionaries rather than mined from relevant domain data, the words in a wrong piece of knowledge usually have a very large probability difference under a topic. However, our knowledge is mined from top words in related topics including topics from the current domain. The words in a piece of incorrect (or correct) knowledge often have similar probabilities under a topic. The proposed dynamic knowledge adjusting mechanism in AKL is superior.

Paired t-test shows that AKL outperforms all baselines significantly ($p < 0.0001$).

6.3 User Evaluation

As our objective is to discover more coherent aspects, we recruited two human judges. Here we also use the test setting 1. Each topic is annotated as coherent if the judge feels that most of its top

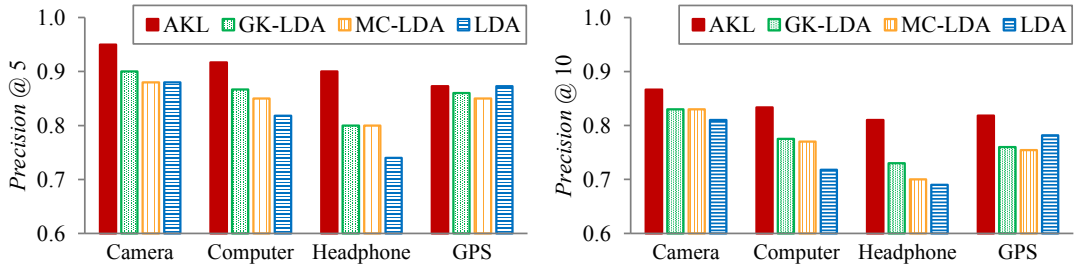

Figure 4: Average $Precision@5$ (Left) and $Precision@10$ (Right) of coherent topics from four models in each domain. (*Headphone* has a lot of overlapping topics in other domains while *GPS* has little.)

terms coherently represent a real-world product aspect; otherwise incoherent. For a coherent topic, each top term is annotated as correct if it reflects the aspect represented by the topic; otherwise incorrect. We labeled the topics of each model at learning iteration 1 where the same pieces of knowledge (extracted from LDA topics at learning iteration 0) are provided to each model. After learning iteration 1, the gap between AKL and the baseline models tends to widen. To be consistent, the results later in Sections 6.4 and 6.5 also show each model at learning iteration 1. We also notice that after a few learning iterations, the topics from AKL model tend to have some resemblance across domains. We found that AKL with 2 learning iterations achieved the best topics. Note that LDA cannot use any prior knowledge.

We manually labeled results from four domains, i.e., Camera, Computer, Headphone, and GPS. We chose Headphone as it has a lot of overlapping of topics with other domains because many electronic products use headphone. GPS was chosen because it does not have much topic overlapping with other domains as its aspects are mostly about Navigation and Maps. Domains Camera and Computer lay in between. We want to see how domain overlapping influences the performance of AKL. Cohen's Kappa scores for annotator agreement are 0.918 (for topics) and 0.872 (for terms).

To measure the results, we compute $Precision@n$ (or $p@n$) based on the annotations, which was also used in (Chen et al., 2013b, Mukherjee and Liu, 2012).

Figure 4 shows the $precision@n$ results for $n = 5$ and 10. We can see that AKL makes improvements in all 4 domains. The improvement varies in domains with the most increase in Headphone and the least in GPS as Headphone overlaps more with other domains than GPS. Note that if a domain shares aspects with many other domains, its model should benefit more; otherwise, it is reasonable to expect lesser improvements. For the baselines, GK-LDA and MC-LDA perform similarly to LDA with minor variations, all of which are inferior to AKL. AKL's improvements over other models are statistically significant based on paired t-test ($p < 0.002$).

In terms of the number of *coherent* topics, AKL discovers one more coherent topic than LDA in Computer and one more coherent topic than GK-LDA and MC-LDA in Headphone. For the other domains, the numbers of coherent topics are the same for all models.

Table 2 shows an example aspect (*battery*) and its top 10 terms produced by AKL and LDA for each domain to give a flavor of the kind of improvements made by AKL. The results for GK-LDA and MC-LDA are about the same as LDA (see also Figure 4). Table 2 focuses on the aspects generated by AKL and LDA. From Table 2, we can see that AKL discovers more correct and meaningful aspect terms at the top. Note that those terms marked in red and italicized are errors. Apart from Table 2, many aspects are dramatically improved by AKL, including some commonly shared aspects such as *Price*, *Screen*, and *Customer Service*.

6.4 Sensitivity to Clustering Parameters

This sub-section investigates the sensitivity of the clustering parameters of AKL (again in test setting 1). The top sub-figure in Figure 5 shows the average Topic Coherence values versus the top k terms per topic used in topic clustering (Section 4.1). The number of clusters is set to the number of domains (see below). We can observe that using $k = 15$ top terms gives the highest value. This is intuitive as too few (or too many) top terms may generate insufficient (or noisy) knowledge.

The bottom sub-figure in Figure 5 shows the average Topic Coherence given different number

Camera		Computer		Headphone		GPS	
AKL	**LDA**	**AKL**	**LDA**	**AKL**	**LDA**	**AKL**	**LDA**
battery	battery	battery	battery	hour	long	battery	*trip*
life	*card*	hour	*cable*	long	battery	hour	battery
hour	*memory*	life	*speaker*	battery	hour	long	hour
long	life	long	*dvi*	life	*comfortable*	*model*	*mile*
charge	usb	*speaker*	*sound*	charge	*easy*	life	long
extra	hour	*sound*	hour	amp	*uncomfortable*	charge	life
minute	minute	charge	*connection*	*uncomfortable*	*headset*	*trip*	*destination*
charger	*sd*	*dvi*	life	*comfortable*	life	*purchase*	*phone*
short	extra	*tv*	*hdmus*	period	*money*	*older*	charge
aa	*device*	*hdmus*	*tv*	*output*	*hard*	*compass*	*mode*

Table 2: Example aspect *Battery* from AKL and LDA in each domain. Errors are italicized in red.

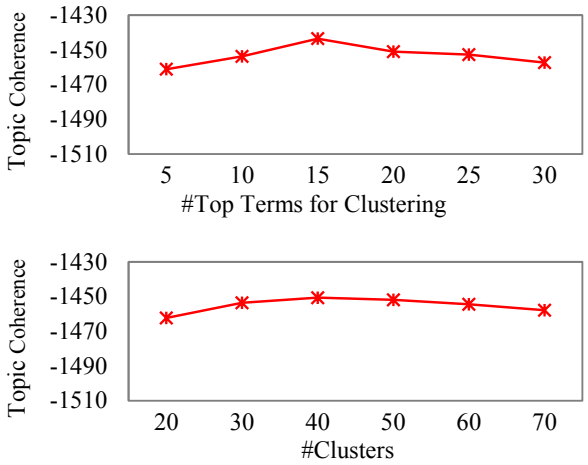

Figure 5: Average topic coherence of AKL versus #top k terms (Top) and #clusters (Bottom).

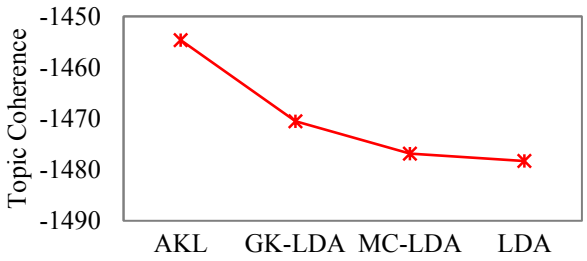

Figure 6: Average topic coherence of each model tested on new/unseen domain.

of clusters. We fix the number of top terms per topic to 15 as it yields the best result (see the top sub-figure in Figure 5). We can see that the performance is not very sensitive to the number of clusters. The model performs similarly for 30 to 50 clusters, with lower Topic Coherence for less than 30 or more than 50 clusters. The significance test indicates that using 30, 40, and 50 clusters, AKL achieved significant improvements over all baseline models ($p < 0.0001$). With more domains, we should expect a larger number of clusters. However, it is difficult to obtain the optimal number of clusters. Thus, we empirically set the number of clusters to the number of domains in our experiments. Note that the number of clusters (C) is expected to be larger than the number of topics in one domain (T) because C is for all domains while T is for one particular domain.

6.5 Test on New Domains

We now evaluate AKL in test setting 2, i.e., the automatically extracted knowledge $\boldsymbol{K}$ (Figure 1) is applied in new/unseen domains other than those in domains $\boldsymbol{D_L}$ used in knowledge learning. The aim is to see how $\boldsymbol{K}$ can help modeling in an unseen domain. In this set of experiments, each domain is tested by using the learned knowledge from the rest 35 domains. Figure 6 shows the average Topic Coherence of each model. The values are also averaged over the 36 tested domains. We can see that AKL achieves the highest Topic Coherence value while LDA has the lowest. The improvements of AKL over all baseline models are significant with $p < 0.0001$.

7 Conclusions

This paper proposed an advanced aspect extraction framework which can learn knowledge automatically from a large number of review corpora and exploit the learned knowledge in extracting more coherent aspects. It first proposed a technique to learn knowledge automatically by clustering and FPM. Then a new topic model with an advanced inference mechanism was proposed to exploit the learned knowledge in a fault-tolerant manner. Experimental results using review corpora from 36 domains showed that the proposed method outperforms state-of-the-art methods significantly.

Acknowledgments

This work was supported in part by a grant from National Science Foundation (NSF) under grant no. IIS-1111092.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In *Proceedings of ICML*, pages 25–32.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of IJCAI*, pages 1171–1177.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A Practical Algorithm for Topic Modeling with Provable Guarantees. In *Proceedings of ICML*, pages 280–288.

David M. Blei and Jon D McAuliffe. 2007. Supervised Topic Models. In *Proceedings of NIPS*, pages 121–128.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

S R K Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2008. Learning Document-Level Semantic Properties from Free-Text Annotations. In *Proceedings of ACL*, pages 263–271.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL*, pages 804–812.

Giuseppe Carenini, Raymond T Ng, and Ed Zwart. 2005. Extracting knowledge from evaluative text. In *Proceedings of K-CAP*, pages 11–18.

Jonathan Chang, Jordan Boyd-Graber, Wang Chong, Sean Gerrish, and David Blei, M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of NIPS*, pages 288–296.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013a. Discovering Coherent Topics Using General Knowledge. In *Proceedings of CIKM*, pages 209–218.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013b. Exploiting Domain Knowledge in Aspect Extraction. In *Proceedings of EMNLP*, pages 1655–1667.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013c. Leveraging Multi-Domain Prior Knowledge in Topic Models. In *Proceedings of IJCAI*, pages 2071–2077.

Yejin Choi and Claire Cardie. 2010. Hierarchical Sequential Learning for Extracting Opinions and their Attributes. In *Proceedings of ACL*, pages 269–274.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of EMNLP*, pages 1277–1287.

Lei Fang and Minlie Huang. 2012. Fine Granular Aspect Analysis using Latent Structural Models. In *Proceedings of ACL*, pages 333–337.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, and Zhong Su. 2009. Product feature categorization with multilevel latent semantic association. In *Proceedings of CIKM*, pages 1087–1096.

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86.

Yulan He, Chenghua Lin, and Harith Alani. 2011. Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification. In *Proceedings of ACL*, pages 123–131.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of UAI*, pages 289–296.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of KDD*, pages 168–177.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive Topic Modeling. In *Proceedings of ACL*, pages 248–257.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of EACL*, pages 204–213.

Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of EMNLP*, pages 1035–1045.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of WSDM*, pages 815–824.

Jeon-hyung Kang, Jun Ma, and Yan Liu. 2012. Transfer Topic Modeling with Ease and Scalability. In *Proceedings of SDM*, pages 564–575.

L Kaufman and P J Rousseeuw. 1990. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. 2013. A Hierarchical Aspect-Sentiment Model for Online Reviews. In *Proceedings of AAAI*, pages 526–533.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining. In *Proceedings of EMNLP*, pages 1065–1074.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of AAAI-CAAW*, pages 100–107.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations. In *Proceedings of ACL*, pages 1630–1639.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Yingju Xia, Shu Zhang, and Hao Yu. 2010. Structure-Aware Review Mining and Summarization. In *Proceedings of COLING*, pages 653–661.

Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011. Generating Aspect-oriented Multi-Document Summarization with Event-aspect model. In *Proceedings of EMNLP*, pages 1137–1146.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM*, pages 375–384.

Kang Liu, Liheng Xu, and Jun Zhao. 2013. Syntactic Patterns versus Word Alignment: Extracting Opinion Targets from Online Reviews. In *Proceedings of ACL*, pages 1754–1763.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Yue Lu and Chengxiang Zhai. 2008. Opinion integration through semi-supervised topic modeling. In *Proceedings of WWW*, pages 121–130.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of WWW*, pages 131–140.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect Sentiment Analysis with Topic Models. In *Proceedings of ICDM Workshops*, pages 81–88.

Yue Lu, Hongning Wang, ChengXiang Zhai, and Dan Roth. 2012. Unsupervised discovery of opposing opinion networks from forum discussions. In *Proceedings of CIKM*, pages 1642–1646.

Hosam Mahmoud. 2008. *Polya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW*, pages 171–180.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of EMNLP*, pages 262–272.

Samaneh Moghaddam and Martin Ester. 2013. The FLDA Model for Aspect-based Opinion Mining: Addressing the Cold Start Problem. In *Proceedings of WWW*, pages 909–918.

Arjun Mukherjee and Bing Liu. 2012. Aspect Extraction through Semi-Supervised Modeling. In *Proceedings of ACL*, pages 339–348.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

James Petterson, Alex Smola, Tibério Caetano, Wray Buntine, and Shravan Narayanamurthy. 2010. Word Features for Latent Dirichlet Allocation. In *Proceedings of NIPS*, pages 1921–1929.

AM Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT*, pages 339–346.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1):9–27.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP*, pages 248–256.

Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. 2010. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38.

Christina Sauper and Regina Barzilay. 2013. Automatic Aggregation by Joint Modeling of Aspects and Values. *J. Artif. Intell. Res. (JAIR)*, 46:89–127.

Swapna Somasundaran and J Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL*, pages 226–234.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL*, pages 308–316.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of KDD*, pages 783–792.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of EMNLP*, pages 1533–1541.

Liheng Xu, Kang Liu, Siwei Lai, Yubo Chen, and Jun Zhao. 2013. Mining Opinion Words and Opinion Targets in a Two-Stage Framework. In *Proceedings of ACL*, pages 1764–1773.

GR Xue, Wenyuan Dai, Q Yang, and Y Yu. 2008. Topic-bridged PLSA for cross-domain text classification. In *Proceedings of SIGIR*, pages 627–634.

Bishan Yang and Claire Cardie. 2013. Joint Inference for Fine-grained Opinion Extraction. In *Proceedings of ACL*, pages 1640–1649.

Shuang Hong Yang, Steven P. Crain, and Hongyuan Zha. 2011. Bridging the language gap: Topic adaptation for documents with different technicality. In *Proceedings of AISTATS*, pages 823–831.

Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews. In *Proceedings of ACL*, pages 1496–1505.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011. Constrained LDA for grouping product features in opinion mining. In *Proceedings of PAKDD*, pages 448–459.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. In *Proceedings of EMNLP*, pages 56–65.

Yanyan Zhao, Bing Qin, and Ting Liu. 2012. Collocation polarity disambiguation using web-based pseudo contexts. In *Proceedings of EMNLP-CoNLL*, pages 160–170.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2013. Collective Opinion Target Extraction in Chinese Microblogs. In *Proceedings of EMNLP*, pages 1840–1850.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of CIKM*, pages 43–50.