Exploiting Domain Knowledge in Aspect Extraction

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

Meichun Hsu, Malu Castellanos, Riddhiman Ghosh
HP Labs
Palo Alto, CA 94304, USA
{meichun.hsu, malu.castellanos, riddhiman.ghosh}@hp.com

Abstract

Aspect extraction is one of the key tasks in sentiment analysis. In recent years, statistical models have been used for the task. However, such models without any domain knowledge often produce aspects that are not interpretable in applications. To tackle the issue, some knowledge-based topic models have been proposed, which allow the user to input some prior domain knowledge to generate coherent aspects. However, existing knowledge-based topic models have several major shortcomings, e.g., little work has been done to incorporate the cannot-link type of knowledge or to automatically adjust the number of topics based on domain knowledge. This paper proposes a more advanced topic model, called *MC-LDA* (LDA with m-set and c-set), to address these problems, which is based on an *Extended generalized Pólya urn* (*E-GPU*) model (which is also proposed in this paper). Experiments on real-life product reviews from a variety of domains show that MC-LDA outperforms the existing state-of-the-art models markedly.

1 Introduction

In sentiment analysis and opinion mining, aspect extraction aims to extract entity aspects or features on which opinions have been expressed (Hu and Liu, 2004; Liu, 2012). For example, in a sentence "The picture looks great," the aspect is "picture." Aspect extraction consists of two sub-tasks: (1) extracting all *aspect terms* (e.g., "picture") from the corpus, and (2) clustering aspect terms with similar meanings (e.g., cluster "picture" and "photo" into one *aspect category* as they mean the same in the domain "Camera"). In this work, we adopt the topic modeling approach as it can perform both sub-tasks simultaneously (see § 2).

Topic models, such as LDA (Blei et al., 2003), provide an unsupervised framework for extracting latent topics in text documents. Topics are aspect categories (or simply aspects) in our context. However, in recent years, researchers have found that fully unsupervised topic models may not produce topics that are very coherent for a particular application. This is because the objective functions of topic models do not always correlate well with human judgments and needs (Chang et al., 2009).

To address the issue, several *knowledge-based topic models* have been proposed. The DF-LDA model (Andrzejewski et al., 2009) incorporates two forms of prior knowledge, also called two types of constraints: *must-links* and *cannot-links*. A must-link states that two words (or terms) should belong to the same topic whereas a cannot-link indicates that two words should not be in the same topic. In (Andrzejewski et al., 2011), more general knowledge can be specified using first-order logic. In (Burns et al., 2012; Jagarlamudi et al., 2012; Lu et al., 2011; Mukherjee and Liu, 2012), seeded models were proposed. They enable the user to specify prior knowledge as seed words/terms for some topics. Petterson et al. (2010) also used word similarity as priors for guidance.

However, none of the existing models is capable of incorporating the cannot-link type of knowledge except DF-LDA (Andrzejewski et al., 2009). Furthermore, none of the existing models, including DF-LDA, is able to automatically adjust the number of topics based on domain knowledge. The domain knowledge, such as cannot-links, may change the number of topics. There are two types of cannot-links: consistent and inconsistent with the domain corpus. For example, in the reviews of

domain "Computer", a topic model may generate two topics *Battery* and *Screen* that represent two different aspects. A cannot-link {battery, screen} as the domain knowledge is thus consistent with the corpus. However, words *Amazon* and *Price* may appear in the same topic due to their high co-occurrences in the Amazon.com review corpus. To separate them, a cannot-link {amazon, price} can be added as the domain knowledge, which is inconsistent with the corpus as these two words have high co-occurrences in the corpus. In this case, the number of topics needs to be *increased* by 1 since the mixed topic has to be separated into two individual topics *Amazon* and *Price*. Apart from the above shortcoming, earlier knowledge-based topic models also have some major shortcomings:

Incapability of handling multiple senses: A word typically has multiple meanings or senses. For example, *light* can mean "of little weight" or "something that makes things visible." DF-LDA cannot handle multiple senses because its definition of must-link is transitive. That is, if A and B form a must-link, and B and C form a must-link, it implies a must-link between A and C, indicating A, B, and C should be in the same topic. This case also applies to the models in (Andrzejewski et al., 2011), (Petterson et al., 2010), and (Mukherjee and Liu, 2012). Although the model in (Jagarlamudi et al., 2012) allows multiple senses, it requires that each topic has at most one set of seed words (seed set), which is restrictive as the amount of knowledge should not be limited.

Sensitivity to the adverse effect of knowledge: When using must-links or seeds, existing models basically try to ensure that the words in a must-link or a seed set have similar probabilities under a topic. This causes a problem: if a must-link comprises of a frequent word and an infrequent word, due to the redistribution of probability mass, the probability of the frequent word will decrease while the probability of the infrequent word will increase. This can harm the final topics because the attenuation of the frequent (often domain important) words can result in some irrelevant words being ranked higher (with higher probabilities).

To address the above shortcomings, we define *m-set* (for *must-set*) as a set of words that should belong to the same topic and *c-set* (*cannot-set*) as a set of words that should not be in the same topic. They are similar to must-link and cannot-link but m-sets do not enforce transitivity. Transitivity is the main cause of the inability to handle multiple senses. Our m-sets and c-sets are also more concise providing knowledge in the context of a set. As in (Andrzejewski et al., 2009), we assume that there is no conflict between m-sets and c-sets, i.e., if w_1 is a cannot-word of w_2 (i.e., shares a c-set with w_2), any word that shares an m-set with w_1 is also a cannot-word of w_2. Note that knowledge as m-sets has also been used in (Chen et al., 2013a) and (Chen et al., 2013b).

We then propose a new topic model, called *MC-LDA* (LDA with m-set and c-set), which is not only able to deal with c-sets and automatically adjust the number of topics, but also deal with the multiple senses and adverse effect of knowledge problems at the same time. For the issue of multiple senses, a new latent variable s is added to LDA to distinguish multiple senses (§ 3). Then, we employ the *generalized Pólya urn* (GPU) model (Mahmoud, 2008) to address the issue of adverse effect of knowledge (§ 4). Deviating from the standard topic modeling approaches, we propose the *Extended generalized Pólya urn* (E-GPU) model (§ 5). E-GPU extends the GPU model to enable multi-urn interactions. This is necessary for handling c-sets and for adjusting the number of topics. E-GPU is the heart of MC-LDA. Due to the extension, a new inference mechanism is designed for MC-LDA (§ 6). Note that E-GPU is generic and can be used in any appropriate application.

In summary, this paper makes the following three contributions:

1. It proposed a new knowledge-based topic model called MC-LDA, which is able to use both m-sets and c-sets, as well as automatically adjust the number of topics based on domain knowledge. At the same time, it can deal with some other major shortcomings of early existing models. To our knowledge, none of the existing knowledge-based models is as comprehensive as MC-LDA in terms of capabilities.
2. It proposed the E-GPU model to enable multi-urn interactions, which enables c-sets to be naturally integrated into a topic model. To the best of our knowledge, E-GPU has not been proposed and used before.
3. A comprehensive evaluation has been conducted to compare MC-LDA with several state-of-the-art models. Experimental results based on both qualitative and quantitative measures demonstrate the superiority of MC-LDA.

2 Related Work

Sentiment analysis has been studied extensively in recent years (Hu and Liu, 2004; Pang and Lee, 2008; Wiebe and Riloff, 2005; Wiebe et al., 2004). According to (Liu, 2012), there are three main approaches to aspect extraction: 1) Using word frequency and syntactic dependency of aspects and sentiment words for extraction (e.g., Blairgoldensohn et al., 2008; Hu and Liu, 2004; Ku et al., 2006; Popescu and Etzioni, 2005; Qiu et al., 2011; Somasundaran and Wiebe, 2009; Wu et al., 2009; Yu et al., 2011; Zhang and Liu, 2011; Zhuang et al., 2006); 2) Using supervised sequence labeling/classification (e.g., Choi and Cardie, 2010; Jakob and Gurevych, 2010; Kobayashi et al., 2007; Li et al., 2010); 3) Topic models (Branavan et al., 2008; Brody and Elhadad, 2010; Fang and Huang, 2012; Jo and Oh, 2011; Kim et al., 2013; Lazaridou et al., 2013; Li et al., 2011; Lin and He, 2009; Lu et al., 2009, 2012, 2011; Lu and Zhai, 2008; Mei et al., 2007; Moghaddam and Ester, 2011; Mukherjee and Liu, 2012; Sauper et al., 2011; Titov and McDonald, 2008; Wang et al., 2010, 2011; Zhao et al., 2010). Other approaches include shallow semantic parsing (Li et al., 2012b), bootstrapping (Xia et al., 2009), Non-English techniques (Abu-Jbara et al., 2013; Zhou et al., 2012), graph-based representation (Wu et al., 2011), convolution kernels (Wiegand and Klakow, 2010) and domain adaption (Li et al., 2012). Stoyanov and Cardie (2011), Wang and Liu (2011), and Meng et al. (2012) studied opinion summarization outside the reviews. Some other works related with sentiment analysis include (Agarwal and Sabharwal, 2012; Kennedy and Inkpen, 2006; Kim et al., 2009; Mohammad et al., 2009).

In this work, we focus on topic models owing to their advantage of performing both aspect extraction and clustering simultaneously. All other approaches only perform extraction. Although there are several related works on clustering aspect terms (e.g., Carenini et al., 2005; Guo et al., 2009; Zhai et al., 2011), they all assume that the aspect terms have been extracted beforehand. We also notice that some aspect extraction models in sentiment analysis separately discover aspect words and aspect specific sentiment words (e.g., Sauper and Barzilay, 2013; Zhao et al., 2010). Our proposed model does not separate them as most sentiment words also imply aspects and most adjectives modify specific attributes of objects. For example, sentiment words *expensive* and *beautiful* imply aspects *price* and *appearance* respectively.

Regarding the knowledge-based models, besides those discussed in § 1, the model (Hu et al., 2011) enables the user to provide guidance interactively. Blei and McAuliffe (2007) and Ramage et al. (2009) used document labels in supervised setting. In (Chen et al., 2013a), we proposed MDK-LDA to leverage multi-domain knowledge, which serves as the basic mechanism to exploit m-sets in MC-LDA. In (Chen et al., 2013b), we proposed a framework (called GK-LDA) to explicitly deal with the wrong knowledge when exploring the lexical semantic relations as the general (domain independent) knowledge in topic models. But these models above did not consider the knowledge in the form of c-sets (or cannot-links).

The *generalized Pólya urn* (GPU) model (Mahmoud, 2008) was first introduced in LDA by Mimno et al. (2011). However, Mimno et al. (2011) did not use domain knowledge. Our results in § 7 show that using domain knowledge can significantly improve aspect extraction. The GPU model was also employed in topic models in our work of (Chen et al., 2013a, 2013b). In this paper, we propose the Extended GPU (E-GPU) model. The E-GPU model is more powerful in handling complex situations in dealing with c-sets.

3 Dealing with M-sets and Multiple Senses

Since the proposed MC-LDA model is a major extension to our earlier work in (Chen et al., 2013a), which can deal with m-sets, we include this earlier work here as the background.

To incorporate m-sets and deal with multiple senses of a word, the MDK-LDA(b) model was proposed in (Chen et al., 2013a), which adds a new latent variable s into LDA. The rationale here is that this new latent variable s guides the model to choose the right sense represented by an m-set. The generative process of MDK-LDA(b) is (the notations are explained in Table 1):

$$\theta \sim Dirichlet(\alpha)$$
$$z_i|\theta_m \sim Multinomial(\theta_m)$$
$$\varphi \sim Dirichlet(\beta)$$
$$s_i|z_i, \varphi \sim Multinomial(\varphi_{z_i})$$
$$\eta \sim Dirichlet(\gamma)$$
$$w_i|z_i, s_i, \eta \sim Multinomial(\eta_{z_i,s_i})$$

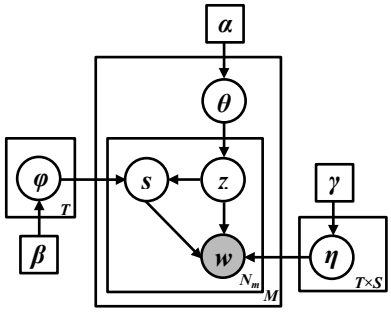

Figure 1. Plate notation for MDK-LDA(b) and MC-LDA.

The corresponding plate is shown in Figure 1. Under MDK-LDA(b), the probability of word w given topic t, i.e., $\pi_t(w)$, is given by:

$$\pi_t(w) = \sum_{s=1}^{S} \varphi_t(s) \cdot \eta_{t,s}(w) \quad (1)$$

where $\varphi_t(s)$ denotes the probability of m-set s occurring under topic t and $\eta_{t,s}(w)$ is the probability of word w appearing in m-set s under topic t.

According to (Chen et al., 2013a), the conditional probability of Gibbs sampler for MDK-LDA(b) is given by (see notations in Table 1):

$$P(z_i = t, s_i = s \,|\, \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma) \propto \frac{n_{m,t}^{-i} + \alpha}{\sum_{t'=1}^{T}(n_{m,t'}^{-i} + \alpha)} \times \frac{n_{t,s}^{-i} + \beta}{\sum_{s'=1}^{S}(n_{t,s'}^{-i} + \beta)} \times \frac{n_{t,s,w_i}^{-i} + \gamma_s}{\sum_{v'=1}^{V}(n_{t,s,v'}^{-i} + \gamma_s)} \quad (2)$$

The superscript $-i$ denotes the counts excluding the current assignments (z_i and s_i) for word w_i.

4 Handling Adverse Effect of Knowledge

4.1 Generalized Pólya urn (GPU) Model

The Pólya urn model involves an urn containing balls of different colors. At discrete time intervals, balls are added or removed from the urn according to their color distributions.

In the simple Pólya urn (SPU) model, a ball is first drawn randomly from the urn and its color is recorded, then that ball is put back along with a new ball of the same color. This selection process is repeated and the contents of the urn change over time, with a self-reinforcing property sometimes expressed as "the rich get richer." SPU is actually exhibited in the Gibbs sampling for LDA.

The *generalized Pólya urn* (GPU) model differs from the SPU model in the replacement scheme during sampling. Specifically, when a ball is randomly drawn, certain numbers of additional balls of each color are returned to the urn, rather than just two balls of the same color as in SPU.

Hyperparameters	
α, β, γ	Dirichlet priors for $\theta,\ \varphi,\ \eta$
Latent & Visible Variables	
z	Topic (Aspect)
s	M-set
w	Word
θ	Document-Topic distribution
θ_m	Topic distribution of document m
φ	Topic-M-set distribution
φ_t	M-set distribution of topic t
η	Topic-M-set-Word distribution
$\eta_{t,s}$	Word distribution of topic t, m-set s
Cardinalities	
M	Number of documents
N_m	Number of words in document m
T	Number of topics
S	Number of m-sets
V	The vocabulary size
Sampling & Count Notations	
z_i	Topic assignment for word w_i
s_i	M-set assignment for word w_i
$\boldsymbol{z}^{-i}$	Topic assignments for all words except w_i
$\boldsymbol{s}^{-i}$	M-set assignments for all words except w_i
$n_{m,t}$	Number of times that topic t is assigned to word tokens in document m
$n_{t,s}$	Number of times that m-set s occurs under topic t
$n_{t,s,w}$	Number of times that word w appears in m-set s under topic t

Table 1. Meanings of symbols.

4.2 Promoting M-sets using GPU

To deal with the issue of sensitivity to the adverse effect of knowledge, MDK-LDA(b) is extended to MDK-LDA which employs the generalized Pólya urn (GPU) sampling scheme**.**

As discussed in § 1, due to the problem of the adverse effect of knowledge, important words may suffer from the presence of rare words in the same m-set. This problem can be dealt with the very sampling scheme of the GPU model (Chen et al., 2013a). Specifically, by adding additional $\mathbb{A}_{s,w',w}$ balls of color s into U_t^S while keeping the drawn ball, we increase the proportion (probability) of seeing the m-set s under topic t and thus promote m-set s as a whole. Consequently, each word in s is more likely to be emitted. We define $\mathbb{A}_{s,w',w}$ as:

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

The corresponding Gibbs sampler for MDK-LDA will be introduced in § 6.

5 Incorporating C-sets

5.1 Extended Generalized Pólya urn Model

To handle the complex situation resulted from incorporating c-sets, we propose an *Extended generalized Pólya urn* (E-GPU) model. Instead of involving only one urn as in SPU and GPU, E-GPU model considers a set of urns in the sampling process. The E-GPU model allows a ball to be transferred from one urn to another, enabling multi-urn interactions. Thus, during sampling, the populations of several urns will evolve even if only one ball is drawn from one urn. This capability makes the E-GPU model more powerful as it models relationships among multiple urns.

We define three sets of urns which will be used in the new sampling scheme in the proposed MC-LDA model. The first set of urns is the topic urns $U^T_{m\in\{1\dots M\}}$, where each topic urn contains T colors (topics) and each ball inside has a color $t \in \{1 \dots T\}$. It corresponds to the document-topic distribution θ in Table 1. The second set of urns (m-set urn $U^S_{t\in\{1\dots T\}}$) corresponds to the topic-m-set distribution φ, with balls of colors (m-sets) $s \in \{1 \dots S\}$ in each m-set urn. The third set of urns is the word urns $U^W_{t,s}$,where $t \in \{1 \dots T\}$ and $s \in \{1 \dots S\}$. Each ball inside a word urn has a color (word) $w \in \{1 \dots V\}$. The distribution η can be reflected in this set of urns.

5.2 Handling C-sets using E-GPU

As MDK-LDA can only use m-sets but not c-sets, we now extend MDK-LDA to the MC-LDA model in order to exploit c-sets. As pointed out in § 1, c-sets may be inconsistent with the corpus domain, which makes them considerably harder to deal with. To tackle the issue, we utilize the proposed E-GPU model and incorporate c-sets handling inside the E-GPU sampling scheme, which is also designed to enable automated adjustment of the number of topics based on domain knowledge.

Based on the definition of c-set, each pair of words in a c-set cannot both have large probabilities under the same topic. As the E-GPU model allows multi-urn interactions, when sampling a ball represents word w from a word urn $U^W_{t,s}$, we want to transfer the balls representing cannot-words of w (sharing a c-set with w) to other urns (see Step 3 a below). That is, decrease the probabilities of those cannot-words under this topic while increasing their corresponding probabilities under some other topics. In order to correctly transfer a ball that represents word w, it should be transferred to an urn which has a higher proportion of w and its related words (i.e., words sharing m-sets with w). That is, we randomly sample an urn that has a higher proportion of any m-set of w to transfer w to (Step 3 b below). However, the situation becomes more involved when a c-set is not consistent with the corpus. For example, aspects *price* and *amazon* may be mixed under one topic (say t) in LDA. The user may want to separate them by providing a c-set {price, amazon}. In this case, according to LDA, word *price* has no topic with a higher proportion of it (and its related words) than topic t. To transfer it, we need to increment the number of topics by 1 and then transfer the word to this new topic urn (step 3 c below). Based on these ideas, we propose the E-GPU sampling scheme for the MC-LDA model below:

1. Sample a topic t from U^T_m, an m-set s from U^S_t, and a word w from $U^W_{t,s}$ sequentially, where m is the mth document.
2. Record t, s and w, put back two balls of color t into urn U^T_m, one ball of color s into urn U^S_t, and two balls of color w into urn $U^W_{t,s}$. Given the matrix $\mathbb{A}$ (in Equation 3), for each word $w' \in s$, we put back $\mathbb{A}_{s,w',w}$ number of balls of color s into urn U^S_t.
3. For each word w_c that shares a c-set with w:
   a) Sample an m-set s_c from U^S_t which satisfies $w_c \in s_c$. Draw a ball b of color w_c (to be transferred) from U^W_{t,s_c} and remove it from U^W_{t,s_c}. The document of ball b is denoted by m_c. If no ball of color w_c can be drawn (i.e., there is no ball of color w_c in U^W_{t,s_c}), skip steps b) to d).
   b) Produce an urn set $\{U^W_{t',s'}\}$ such that each urn in it satisfies the following conditions:
      i) $t' \neq t, w_c \in s'$
      ii) The proportion of balls of color s' in $U^S_{t'}$ is higher than that of balls of color s_c in U^S_t.
   c) If $\{U^W_{t',s'}\}$ is not empty, randomly select one urn $U^W_{t',s'}$ from it. If $\{U^W_{t',s'}\}$ is empty, set $T = T + 1, t' = T$, draw an m-set s' from $U^S_{t'}$ which satisfies $w_c \in s'$. Record s' for step d).
   d) Put the ball b drawn from Step a) into $U^W_{t',s'}$, as well as a ball of color s' into $U^S_{t'}$ and a ball of color t' into $U^T_{m_c}$.

Note that the E-GPU model cannot be reflected in the graphical model in Figure 1 as it is essentially

Algorithm 1. GibbsSampling(m, w_i, $\mathbb{\Delta}$, μ, Ω)

Input: Document m, Word w_i, Matrix $\mathbb{\Delta}$,
Transfer cannot-word flag μ,
A set of valid topics Ω to be assigned to w_i

1: $n_{m,z_i} \leftarrow n_{m,z_i} - 1$;
2: **for** each word w' in s_i **do**
3: $n_{z_i,s_i} \leftarrow n_{z_i,s_i} - \mathbb{\Delta}_{s_i,w',w_i}$;
4: **end for**
5: $n_{z_i,s_i,w_i} \leftarrow n_{z_i,s_i,w_i} - 1$;
6: Jointly sample $z_i \in \Omega$ and $s_i \ni w_i$ using Equation 2;
7: $n_{m,z_i} \leftarrow n_{m,z_i} + 1$;
8: **for** each word w' in s_i **do**
9: $n_{z_i,s_i} \leftarrow n_{z_i,s_i} + \mathbb{\Delta}_{s_i,w',w_i}$;
10: **end for**
11: $n_{z_i,s_i,w_i} \leftarrow n_{z_i,s_i,w_i} + 1$;
12: **if** μ is true **then**
13: TransferCannotWords(w_i, z_i);
14: **end if**

Figure 2. Gibbs sampling for MC-LDA.

sampling scheme, and hence MC-LDA shares the same plate as MDK-LDA(b).

6 Collapsed Gibbs Sampling

We now describe the collapsed Gibbs sampler (Griffiths and Steyvers, 2004) with the detailed conditional distributions and algorithms for MC-LDA. Inference of z and s can be computationally expensive due to the non-exchangeability of words under the E-GPU models. We take the approach of (Mimno et al., 2011) which approximates the true Gibbs sampling distribution by treating each word as if it were the last.

For each word w_i, we perform hierarchical sampling consisting of the following three steps (the detailed algorithms are given in Figures 2 and 3):

Step 1 (Lines 1-11 in Figure 2): We jointly sample a topic z_i and an m-set s_i (containing w_i) for w_i, which gives us a blocked Gibbs sampler (Ishwaran and James, 2001), with the conditional probability given by:

$$P(z_i = t, s_i = s | \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma, \mathbb{\Delta}) \propto \frac{n_{m,t}^{-i} + \alpha}{\sum_{t'=1}^{T}\left(n_{m,t'}^{-i} + \alpha\right)} \times \frac{\sum_{w'=1}^{V}\sum_{v'=1}^{V} \mathbb{\Delta}_{s,v',w'} \cdot n_{t,s,v'}^{-i} + \beta}{\sum_{s'=1}^{S}\left(\sum_{w'=1}^{V}\sum_{v'=1}^{V} \mathbb{\Delta}_{s',v',w'} \cdot n_{t,s',v'}^{-i} + \beta\right)} \times \frac{n_{t,s,w_i}^{-i} + \gamma_s}{\sum_{v'=1}^{V}\left(n_{t,s,v'}^{-i} + \gamma_s\right)} \quad (4)$$

This step is the same as the Gibbs sampling for the MDK-LDA model.

Algorithm 2.TransferCannotWords(w_i, z_i)

Input: Word w_i, Topic z_i,

1: **for** each cannot-word w_c of w_i **do**
2: Randomly select an m-set s_c from all m-sets of w_c;
3: Build a set Ψ containing all the instances of w_c from the corpus with topic and m-set assignments being z_i and s_c;
4: **if** Ψ is not empty **then**
5: Draw an instance of w_c from Ψ (denoting the document of this instance by m_c) using Equation 5;
6: Generate a topic set Ω' that each topic t' inside satisfies $max_{s' \ni w_c}(\varphi_{t'}(s')) > \varphi_{z_i}(s_c)$.
7: **if** Ω' is not empty **then**
8: GibbsSampling(m_c, w_c, $\mathbb{\Delta}$, false, Ω');
9: **else**
10: $Dummy = T + 1$; // T is #Topics.
11: GibbsSampling(m_c, w_c, $\mathbb{\Delta}$, false, $\{Dummy\}$);
12: **end if**
13: **end if**
14: **end for**

Figure 3. Transfer cannot-words in Gibbs sampling.

Step 2 (lines 1-5 in Figure 3): For every cannot-word (say w_c) of w_i, randomly pick an urn U_{z_i,s_c}^{W} from the urn set $\{U_{z_i,\bar{s}}^{W}\}$ where $\bar{s} \ni w_c$. If there exists at least one ball of color w_c in urn U_{z_i,s_c}^{W}, we sample one ball (say b_c) of color w_c from urn U_{z_i,s_c}^{W}, based on the following conditional distribution:

$$P(b = b_c | \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}, \alpha, \beta, \gamma, \mathbb{\Delta}) \propto \frac{n_{m_c,t} + \alpha}{\sum_{t'=1}^{T}\left(n_{m_c,t'} + \alpha\right)} \quad (5)$$

where m_c denotes the document of the ball b_c of color w_c.

Step 3 (lines 6-12 in Figure 3): For each drawn ball b from Step 2, resample a topic t and an m-set s (containing w_c) based on the following conditional distribution:

$$\begin{aligned} & P(z_b = t, s_b = s | \boldsymbol{z}^{-b}, \boldsymbol{s}^{-b}, \boldsymbol{w}, \alpha, \beta, \gamma, \mathbb{\Delta}, b = b_c) \\ & \propto \mathbf{I}_{\left[0, \max\limits_{s' \ni w_c}(\varphi_t(s'))\right]}\left(\varphi_{z_c}(s_c)\right) \times \frac{n_{m,t}^{-b} + \alpha}{\sum_{t'=1}^{T}\left(n_{m,t'}^{-b} + \alpha\right)} \\ & \times \frac{\sum_{w'=1}^{V}\sum_{v'=1}^{V} \mathbb{\Delta}_{s,v',w'} \cdot n_{t,s,v'}^{-b} + \beta}{\sum_{s'=1}^{S}\left(\sum_{w'=1}^{V}\sum_{v'=1}^{V} \mathbb{\Delta}_{s',v',w'} \cdot n_{t,s',v'}^{-b} + \beta\right)} \\ & \times \frac{n_{t,s,w_b}^{-b} + \gamma_s}{\sum_{v'=1}^{V}\left(n_{t,s,v'}^{-b} + \gamma_s\right)} \end{aligned} \quad (6)$$

where z_c (same as z_i in Figure 3) and s_c are the original topic and m-set assignments. The superscript $-b$ denotes the counts excluding the original

assignments. **I**() is an indicator function, which restricts the ball to be transferred only to an urn that contains a higher proportion of its m-set. When no topic t can be successfully sampled and the current sweep (iteration) of Gibbs sampling has the same number of topic (T) as the previous sweep, we increment T by 1. And then assign T to z_b. The counts and parameters are also updated accordingly.

7 Experiments

We now evaluate the proposed MC-LDA model and compare it with state-of-the-art existing models. Two unsupervised baseline models that we compare with are:

- **LDA**: LDA is the basic unsupervised topic model (Blei et al., 2003).
- **LDA-GPU**: LDA with GPU (Mimno et al., 2011). Specifically, LDA-GPU applies GPU in LDA using co-document frequency.

As for knowledge-based models, we focus on comparing with DF-LDA model (Andrzejewski et al., 2009), which is perhaps the best known knowledge-based model and it allows both must-links and cannot-links.

For a comprehensive evaluation, we consider the following variations of MC-LDA and DF-LDA:

- **MC-LDA:** MC-LDA with both m-sets and c-sets. This is the newly proposed model.
- **M-LDA**: MC-LDA with m-sets only. This is the MDK-LDA model in (Chen et al., 2013a).
- **DF-M**: DF-LDA with must-links only.
- **DF-MC**: DF-LDA with both must-links and cannot-links. This is the full DF-LDA model in (Andrzejewski et al., 2009).

We do not compare with seeded models in (Burns et al., 2012; Jagarlamudi et al., 2012; Lu et al., 2011; Mukherjee and Liu, 2012) as seed sets are special cases of must-links and they also do not allow c-sets (or cannot-links).

7.1 Datasets and Settings

Datasets: We use product reviews from four domains (types of products) from Amazon.com for evaluation. The corpus statistics are shown in Table 2 (columns 2 and 3). The domains are "Camera," "Food," "Computer," and "Care" (short for "Personal Care"). We have made the datasets publically available at the website of the first author.

Domain	#Reviews	#Sentences	#M-sets	#C-sets
Camera	500	5171	173	18
Food	500	2416	85	10
Computer	500	2864	92	6
Care	500	3008	119	13
Average	500	3116	103	9

Table 2. Corpus statistics with #m-sets and #c-sets having at least two words.

Pre-processing: We ran the Stanford Core NLP Tools[1] to perform sentence detection and lemmatization. Punctuations, stopwords[2], numbers and words appearing less than 5 times in each corpus were removed. The domain name was also removed, e.g., word *camera* in the domain "Camera", since it co-occurs with most words in the corpus, leading to high similarity among topics/aspects.

Sentences as documents: As noted in (Titov and McDonald, 2008), when standard topic models are applied to reviews as documents, they tend to produce topics that correspond to global properties of products (e.g., brand name), which make topics overlapping with each other. The reason is that all reviews of the same type of products discuss about the same aspects of these products. Only the brand names and product names are different. Thus, using individual reviews for modeling is not very effective. Although there are approaches which model sentences (Jo and Oh, 2011; Titov and McDonald, 2008), we take the approach of (Brody and Elhadad, 2010), dividing each review into sentences and treating each sentence as an independent document. Sentences can be used by all three baselines without any change to their models. Although the relationships between sentences are lost, the data is fair to all models.

Parameter settings: For all models, posterior inference was drawn using 1000 Gibbs iterations with an initial burn-in of 100 iterations. For all models, we set $\alpha = 1$ and $\beta = 0.1$. We found that small changes of α and β did not affect the results much, which was also reported in (Jo and Oh, 2011) who also used online reviews. For the number of topics T, we tried different values (see §7.2) as it is hard to know the exact number of topics. While non-parametric Bayesian approaches (Teh et al., 2006) aim to estimate T from the corpus, they are often sensitive to the hyper-parameters (Heinrich, 2009).

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list

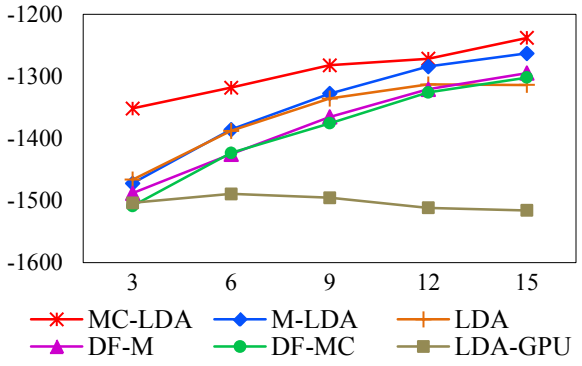

Figure 4. Avg. Topic Coherence score of each model across different number of topics.

For DF-LDA, we followed (Andrzejewski et al., 2009) to generate must-links and cannot-links from our domain knowledge. We then ran DF-LDA[3] while keeping its parameters as proposed in (Andrzejewski et al., 2009) (we also experimented with different parameter settings but they did not produce better results). For our proposed model, we estimated the thresholds using cross validation in our pilot experiments. Estimated value $\sigma = 0.2$ in equation 3 yielded good results. The second stage (steps 2 and 3) of the Gibbs sampler for MC-LDA (for dealing with c-sets) is applied after burn-in phrase.

Domain knowledge: User knowledge about a domain can vary a great deal. Different users may have very different knowledge. To reduce this variance for a more reliable evaluation, instead of asking a human user to provide m-sets, we obtain the synonym sets and the antonym sets of each word that is a noun or adjective (as words of other parts-of-speech usually do not indicate aspects) from WordNet (Miller, 1995) and manually verify the words in those sets for the domain. Note that if a word w is not provided with any m-set, it is treated as a singleton m-set $\{w\}$. For c-sets, we ran LDA in each domain and provide c-sets based on the wrong results of LDA as in (Andrzejewski et al., 2009). Then, the knowledge is provided to each model in the format required by each model. The numbers of m-sets and c-sets are listed in columns 4 and 5 of Table 2. Duplicate sets have been removed.

7.2 Objective Evaluation

In this section, we evaluate our proposed MC-LDA model objectively. Topic models are often evaluated using perplexity on held-out test data. However, the perplexity metric does not reflect the semantic coherence of individual topics learned by a topic model (Newman et al., 2010). Recent research has shown potential issues with perplexity as a measure: (Chang et al., 2009) suggested that the perplexity can sometimes be contrary to human judgments. Also, perplexity does not really reflect our goal of finding coherent aspects with accurate semantic clustering. It only provides a measure of how well the model fits the data.

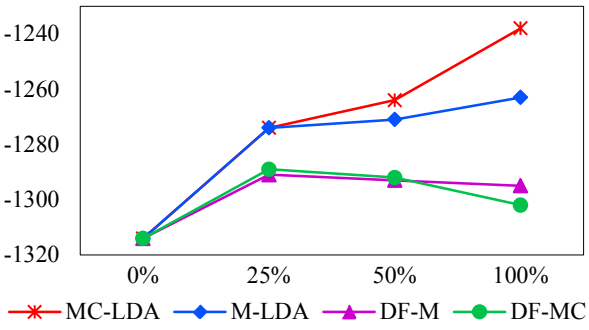

Figure 5. Avg. Topic Coherence score for different proportions of knowledge.

The *Topic Coherence* metric (Mimno et al., 2011) (also called the "UMass" measure (Stevens and Buttler, 2012)) was proposed as a better alternative for assessing topic quality. This metric relies upon word co-occurrence statistics within the documents, and does not depend on external resources or human labeling. It was shown that topic coherence is highly consistent with human expert labeling by Mimno et al. (2011). Higher topic coherence score indicates higher quality of topics, i.e., better topic interpretability.

Effects of Number of Topics

Since our proposed models and the baseline models are all parametric models, we first compare each model given different numbers of topics. Figure 4 shows the average Topic Coherence score of each model given different numbers of topics. From Figure 4, we note the following:

1. MC-LDA consistently achieves the highest Topic Coherence scores given different numbers of topics. M-LDA also works better than the other baseline models, but not as well as MC-LDA. This shows that both m-sets and c-sets are beneficial in producing coherent aspects.
2. DF-LDA variants, DF-M and DF-MC, do not perform well due to the shortcomings discussed

[3] http://pages.cs.wisc.edu/~andrzeje/research/df_lda.html

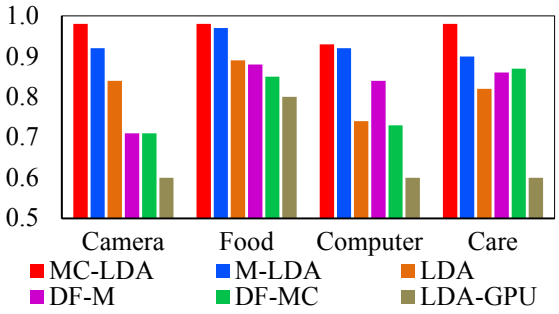

Figure 6. Avg. *p@5* of good topics for each model across different domains.

The models of each bar from left to rights are MC-LDA, M-LDA, LDA, DF-M, DF-MC, LDA-GPU. (Same for Figure 7)

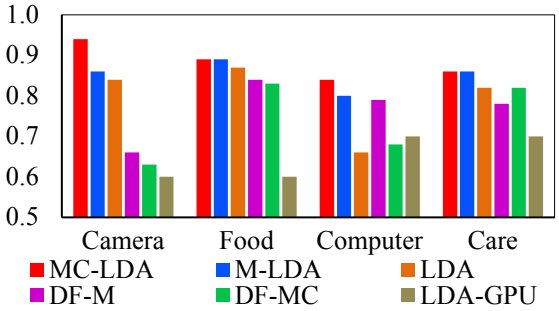

Figure 7. Avg. *p@10* of good topics for each model across different domains.

in § 1. It is slightly better than LDA when $T = 15$, but worse than LDA in other cases. We will further analyze the effects of knowledge on MC-LDA and DF-LDA shortly.

3. LDA-GPU does not perform well due to its use of co-document frequency. As frequent words usually have high co-document frequency with many other words, the frequent words are ranked top in many topics. This shows that the guidance using domain knowledge is more effective than using co-document frequency.

In terms of improvements, MC-LDA outperforms M-LDA significantly ($p < 0.03$) and all other baseline models significantly ($p < 0.01$) based on a paired *t*-test. It is important to note that by no means do we say that LDA-GPU and DF-LDA are not effective. We only say that for the task of aspect extraction and leveraging domain knowledge, these models do not generate as coherent aspects as ours because of their shortcomings discussed in § 1. In general, with more topics, the Topic Coherence scores increase. We found that when T is larger than 15, aspects found by each model became more and more overlapping, with several aspects expressing the same features of products. So we fix $T = 15$ in the subsequent experiments.

Effects of Knowledge

To further analyze the effects of knowledge on models, in each domain, we randomly sampled different proportions of knowledge (i.e., different numbers of m-sets/must-links and c-sets/cannot-links) as shown in Figure 5, where 0% means no knowledge (same as LDA and LDA-GPU, which do not incorporate knowledge) and 100% means all knowledge. From Figure 5, we see that MC-LDA and M-LDA both perform consistently better than DF-MC and DF-M across different proportions of knowledge. With the increasing number of knowledge sets, MC-LDA and M-LDA achieve higher Topic Coherence scores (i.e., produce more coherent aspects). In general, MC-LDA performs the best. For both DF-MC and DF-M, the Topic Coherence score increases from 0% to 25% knowledge, but decreases with more knowledge (50% and 100%). This shows that with limited amount of knowledge, the shortcomings of DF-LDA are not very obvious, but with more knowledge, these issues become more serious and thus degrade the performance of DF-LDA.

7.3 Human Evaluation

Since our aim is to make topics more interpretable and conformable to human judgments, we worked with two judges who are familiar with Amazon products and reviews to evaluate the models subjectively. Since topics from topic models are rankings based on word probability and we do not know the number of correct topical words, a natural way to evaluate these rankings is to use *Precision@n* (or *p@n*) which was also used in (Mukherjee and Liu, 2012; Zhao et al., 2010), where *n* is the rank position. We give *p@n* for *n* = 5 and 10. There are two steps in human evaluation: topic labeling and word labeling.

Topic Labeling: We followed the instructions in (Mimno et al., 2011) and asked the judges to label each topic as *good* or *bad*. Each topic was presented as a list of 10 most probable words in descending order of their probabilities under that topic. The models which generated the topics for labeling were obscure to the judges. In general, each topic was annotated as *good* if it had more than half of its words coherently related to each other representing a semantic concept together; otherwise *bad*. Agreement of human judges on topic

labeling using Cohen's Kappa yielded a score of 0.92 indicating almost perfect agreements according to the scale in (Landis and Koch, 1977). This is reasonable as topic labeling is an easy task and semantic coherence can be judged well by humans.

Word Labeling: After topic labeling, we chose the topics, which were labeled as good by both judges, as good topics. Then, we asked the two judges to label each word of the top 10 words in these good topics. Each word was annotated as *correct* if it was coherently related to the concept represented by the topic; otherwise *incorrect*. Since judges already had the conception of each topic in mind when they were labeling topics, labeling each word was not difficult which explains the high Kappa score for this labeling task (score = 0.892).

Quantitative Results

Figures 6 and 7 give the average *p@5* and *p@10* of all good topics over all four domains. The numbers of good topics generated by each model are shown in Table 3. We can see that the human evaluation results are highly consistent with Topic Coherence results in §7.2. MC-LDA improves over M-LDA significantly ($p < 0.01$) and both MC-LDA and M-LDA outperforms the other baseline models significantly ($p < 0.005$) using a paired *t*-test. We also found that when the domain knowledge is simple with one word usually expressing only one meaning/sense (e.g., in the domain "Computer"), DF-LDA performs better than LDA. In other domains, it performs similarly or worse than LDA. Again, it shows that DF-LDA is not effective to handle complex knowledge, which is consistent with the results of effects of knowledge on DF-LDA in §7.2.

Qualitative Results

We now show some qualitative results to give an intuitive feeling of the outputs from different models. There are a large number of aspects that are dramatically improved by MC-LDA. Due to space constraints, we only show some examples. To further focus, we just show some results of MC-LDA, M-LDA and LDA. The results from LDA-GPU and DF-LDA were inferior and hard for the human judges to match them with aspects found by the other models for qualitative comparison.

Table 4 shows three aspects Amazon, Price, Battery generated by each model in the domain

#Good Topics	MC-LDA	M-LDA	LDA	DF-M	DF-MC	LDA-GPU
Camera	**15/18**	12	11	9	7	3
Food	**8/16**	7	7	5	4	5
Computer	**12/16**	10	7	9	6	4
Care	**11/16**	10	9	10	9	3
Average	**11.5/16.5**	9.75/15	8.5/15	8.25/15	6.5/15	3.75/15

Table 3. Number of good topics of each model.

In x/y, x is the number of discovered good topics, and y is the total number of topics generated.

MC-LDA			M-LDA		LDA	
Amazon	**Price**	**Battery**	**Price**	**Battery**	**Amazon**	**Battery**
review	price	battery	price	battery	*card*	battery
amazon	*perform*	life	*lot*	*review*	*day*	*screen*
software	money	*day*	money	*amazon*	amazon	life
customer	expensive	extra	*big*	life	*memory*	*lcd*
month	cost	charger	expensive	extra	product	*water*
support	*week*	*water*	*point*	*day*	*sd*	usb
warranty	cheap	time	cost	power	*week*	*cable*
package	purchase	power	*photo*	time	*month*	*case*
product	deal	hour	*dot*	*support*	item	charger
hardware	product	aa	purchase	*customer*	*class*	hour

Table 4. Example aspects in the domain "Camera"; errors are marked in red/italic.

"Camera". Both LDA and M-LDA can only discover two aspects but M-LDA has a higher average precision. Given the c-set {amazon, price, battery}, MC-LDA can discover all three aspects with the highest average precision.

8 Conclusion

This paper proposed a new model to exploit domain knowledge in the form of m-sets and c-sets to generate coherent aspects (topics) from online reviews. The paper first identified and characterized some shortcomings of the existing knowledge-based models. A new model called MC-LDA was then proposed, whose sampling scheme was based on the proposed Extended GPU (E-GPU) model enabling multi-urn interactions. A comprehensive evaluation using real-life online reviews from multiple domains shows that MC-LDA outperforms the state-of-the-art models significantly and discovers aspects with high semantic coherence. In our future work, we plan to incorporate aspect specific sentiments in the MC-LDA model.

Acknowledgments

This work was supported in part by a grant from National Science Foundation (NSF) under grant no. IIS-1111092, and a grant from HP Labs Innovation Research Program.

References

Amjad Abu-Jbara, Ben King, Mona Diab, and Dragomir Radev. 2013. Identifying Opinion Subgroups in Arabic Online Discussions. In *Proceedings of ACL*.

Apoorv Agarwal and Jasneet Sabharwal. 2012. End-to-End Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, at the 24th International Conference on Computational Linguistics (IEEASMD-COLING 2012)*, Vol. 2.

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.

Sasha Blair-goldensohn, Tyler Neylon, Kerry Hannan, George A. Reis, Ryan Mcdonald, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *Proceedings of In NLP in the Information Explosion Era*.

David M. Blei and Jon D. McAuliffe. 2007. Supervised Topic Models. In *Proceedings of NIPS*.

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.

Nicola Burns, Yaxin Bi, Hui Wang, and Terry Anderson. 2012. Extended Twofold-LDA Model for Two Aspects in One Sentence. *Advances in Computational Intelligence*, Vol. 298, pages 265–275. Springer Berlin Heidelberg.

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. Leveraging Multi-Domain Prior Knowledge in Topic Models. In *Proceedings of IJCAI*, pages 2071–2077.

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

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

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

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *PNAS*, *101 Suppl*, 5228–5235.

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.

Gregor Heinrich. 2009. A Generic Approach to Topic Models. In *Proceedings of ECML PKDD*, pages 517 – 532.

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.

Hemant Ishwaran and LF James. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

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.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, *22*(2), 110–125.

Jungi Kim, Jinji Li, and Jong-Hyeok Lee. 2009. Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis. In *Proceedings of ACL/IJCNLP*, pages 253–261.

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

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 Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 100–107.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, *33*.

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*.

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.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012a. Cross-Domain Co-Extraction of Sentiment and Topic Lexicons. In *Proceedings of ACL (1)*, pages 410–419.

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.

Shoushan Li, Rongyang Wang, and Guodong Zhou. 2012b. Opinion Target Extraction Using a Shallow Semantic Parsing Framework. In *Proceedings of AAAI*.

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

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

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.

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.

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.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. 2012. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of KDD*, pages 379–387.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, *38*(11), 39–41.

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. 2011. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *Proceedings of SIGIR*, pages 665–674.

Saif Mohammad, Cody Dunne, and Bonnie J. Dorr. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. In *Proceedings of EMNLP*, pages 599–608.

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

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of JCDL*, pages 215–224.

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.

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

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2011. Content Models with Attitude. In *Proceedings of ACL*, pages 350–358.

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

Keith Stevens and PKDAD Buttler. 2012. Exploring Topic Coherence over many models and many topics. In *Proceedings of EMNLP-CoNLL*, pages 952–961.

Veselin Stoyanov and Claire Cardie. 2011. Automatically Creating General-Purpose Opinion

Summaries from Text. In *Proceedings of RANLP*, pages 202–209.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 1–30.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW*, pages 111–120.

Dong Wang and Yang Liu. 2011. A Pilot Study of Opinion Summarization in Conversations. In *Proceedings of ACL*, pages 331–339.

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.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of KDD*, pages 618–626.

Janyce Wiebe and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceedings of CICLing*, pages 486–497.

Janyce Wiebe, Theresa Wilson, Rebecca F. Bruce, Matthew Bell, and Melanie Martin. 2004. Learning Subjective Language. *Computational Linguistics*, *30*(3), 277–308.

Michael Wiegand and Dietrich Klakow. 2010. Convolution Kernels for Opinion Holder Extraction. In *Proceedings of HLT-NAACL*, pages 795–803.

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

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2011. Structural Opinion Mining for Graph-based Sentiment Representation. In *Proceedings of EMNLP*, pages 1332–1341.

Yunqing Xia, Boyi Hao, and Kam-Fai Wong. 2009. Opinion Target Network and Bootstrapping Method for Chinese Opinion Target Extraction. In *Proceedings of AIRS*, pages 339–350.

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 the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 448–459.

Lei Zhang and Bing Liu. 2011. Identifying Noun Product Features that Imply Opinions. In *Proceedings of ACL (Short Papers)*, pages 575–580.

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.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2012. Cross-Language Opinion Target Extraction in Review Texts. In *Proceedings of ICDM*, pages 1200–1205.

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