Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid

Wayne Xin Zhao[†], **Jing Jiang**[‡], **Hongfei Yan**[†], **Xiaoming Li**[†]
[†]School of Electronics Engineering and Computer Science, Peking University, China
[‡]School of Information Systems, Singapore Management University, Singapore
{zhaoxin,yhf}@net.pku.edu.cn, jingjiang@smu.edu.cn, lxm@pku.edu.cn

Abstract

Discovering and summarizing opinions from online reviews is an important and challenging task. A commonly-adopted framework generates structured review summaries with aspects and opinions. Recently topic models have been used to identify meaningful review aspects, but existing topic models do not identify aspect-specific opinion words. In this paper, we propose a MaxEnt-LDA hybrid model to jointly discover both aspects and aspect-specific opinion words. We show that with a relatively small amount of training data, our model can effectively identify aspect and opinion words simultaneously. We also demonstrate the domain adaptability of our model.

1 Introduction

With the dramatic growth of opinionated user-generated content, consumers often turn to online product reviews to seek advice while companies see reviews as a valuable source of consumer feedback. How to automatically understand, extract and summarize the opinions expressed in online reviews has therefore become an important research topic and gained much attention in recent years (Pang and Lee, 2008). A wide spectrum of tasks have been studied under review mining, ranging from coarse-grained document-level polarity classification (Pang et al., 2002) to fine-grained extraction of opinion expressions and their targets (Wu et al., 2009). In particular, a general framework of summarizing reviews of a certain product is to first identify different aspects (a.k.a. features) of the given product and then extract specific opinion expressions for each aspect. For example, aspects of a restaurant may include *food*, *staff*, *ambience* and *price*, and opinion expressions for *staff* may include *friendly*, *rude*, etc. Because of the practicality of this structured summary format, it has been adopted in several previous studies (Hu and Liu, 2004; Popescu and Etzioni, 2005; Brody and Elhadad, 2010) as well as some commercial systems, e.g. the "scorecard" feature at Bing shopping[1].

Different approaches have been proposed to identify aspect words and phrases from reviews. Previous methods using frequent itemset mining (Hu and Liu, 2004) or supervised learning (Jin and Ho, 2009; Jin et al., 2009; Wu et al., 2009) have the limitation that they do not group semantically related aspect expressions together. Supervised learning also suffers from its heavy dependence on training data. In contrast, unsupervised, knowledge-lean topic modeling approach has been shown to be effective in automatically identifying aspects and their representative words (Titov and McDonald, 2008; Brody and Elhadad, 2010). For example, words such as *waiter*, *waitress*, *staff* and *service* are grouped into one aspect.

We follow this promising direction and extend existing topic models to jointly identify both aspect and opinion words, especially aspect-specific opinion words. Current topic models for opinion mining, which we will review in detail in Section 2, still lack this ability. But separating aspect and opinion words can be very useful. Aspect-specific opinion words can be used to construct a domain-dependent senti-

[1] http://www.bing.com/shopping

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 56–65,
MIT, Massachusetts, USA, 9-11 October 2010.

ment lexicon and applied to tasks such as sentiment classification. They can also provide more informative descriptions of the product or service being reviewed. For example, using more specific opinion words such as *cozy* and *romantic* to describe the *ambience* aspect in a review summary is more meaningful than using generic words such as *nice* and *great*. To the best of our knowledge, Brody and Elhadad (2010) are the first to study aspect-specific opinion words, but their opinion word detection is performed outside of topic modeling, and they only consider adjectives as possible opinion words.

In this paper, we propose a new topic modeling approach that can automatically separate aspect and opinion words. A novelty of this model is the integration of a discriminative maximum entropy (MaxEnt) component with the standard generative component. The MaxEnt component allows us to leverage arbitrary features such as POS tags to help separate aspect and opinion words. Because the supervision relies mostly on non-lexical features, although our model is no longer fully unsupervised, the number of training sentences needed is relatively small. Moreover, training data can also come from a different domain and yet still remain effective, making our model highly domain adaptive. Empirical evaluation on large review data sets shows that our model can effectively identify both aspects and aspect-specific opinion words with a small amount of training data.

2 Related Work

Pioneered by the work of Hu and Liu (2004), review summarization has been an important research topic. There are usually two major tasks involved, namely, aspect or feature identification and opinion extraction. Hu and Liu (2004) applied frequent itemset mining to identify product features without supervision, and considered adjectives collocated with feature words as opinion words. Jin and Ho (2009), Jin et al. (2009) and Wu et al. (2009) used supervised learning that requires hand-labeled training sentences to identify both aspects and opinions. A common limitation of these methods is that they do not group semantically related aspect expressions together. Furthermore, supervised learning usually requires a large amount of training data in order to perform well and is not easily domain adaptable.

Topic modeling provides an unsupervised and knowledge-lean approach to opinion mining. Titov and McDonald (2008) show that global topic models such as LDA (Blei et al., 2003) may not be suitable for detecting rateable aspects. They propose multi-grain topic models for discovering local rateable aspects. However, they do not explicitly separate aspect and opinion words. Lin and He (2009) propose a joint topic-sentiment model, but topic words and sentiment words are still not explicitly separated. Mei et al. (2007) propose to separate topic and sentiment words using a positive sentiment model and a negative sentiment model, but both models capture general opinion words only. In contrast, we model *aspect-specific* opinion words as well as general opinion words.

Recently Brody and Elhadad (2010) propose to detect aspect-specific opinion words in an unsupervised manner. They take a two-step approach by first detecting aspect words using topic models and then identifying aspect-specific opinion words using polarity propagation. They only consider adjectives as opinion words, which may potentially miss opinion words with other POS tags. We try to jointly capture both aspect and opinion words within topic models, and we allow non-adjective opinion words.

Another line of related work is about how to incorporate useful features into topic models (Zhu and Xing, 2010; Mimno and McCallum, 2008). Our MaxEnt-LDA hybrid bears similarity to these recent models but ours is designed for opinion mining.

3 Model Description

Our model is an extension of LDA (Blei et al., 2003) but captures both aspect words and opinion words. To model the aspect words, we use a modified version of the multi-grain topic models from (Titov and McDonald, 2008). Our model is simpler and yet still produces meaningful aspects. Specifically, we assume that there are T aspects in a given collection of reviews from the same domain, and each review document contains a mixture of aspects. We further assume that each sentence (instead of each word as in standard LDA) is assigned to a single aspect, which is often true based on our observation.

To understand how we model the opinion words, let us first look at two example review sentences

from the restaurant domain:

The food was tasty.

The waiter was quite friendly.

We can see that there is a strong association of *tasty* with *food* and similarly of *friendly* with *waiter*. While both *tasty* and *friendly* are specific to the restaurant domain, they are each associated with only a single aspect, namely *food* and *staff*, respectively. Besides these aspect-specific opinion words, we also see general opinion words such as *great* in the sentence "*The food was great!*" These general opinion words are shared across aspects, as opposed to aspect-specific opinion words which are used most commonly with their corresponding aspects. We therefore introduce a general opinion model and T aspect-specific opinion models to capture these different opinion words.

3.1 Generative Process

We now describe the generative process of the model. First, we draw several multinomial word distributions from a symmetric Dirichlet prior with parameter β: a background model $\phi^{\mathcal{B}}$, a general aspect model $\phi^{\mathcal{A},g}$, a general opinion model $\phi^{\mathcal{O},g}$, T aspect models $\{\phi^{\mathcal{A},t}\}_{t=1}^{T}$ and T aspect-specific opinion models $\{\phi^{\mathcal{O},t}\}_{t=1}^{T}$. All these are multinomial distributions over the vocabulary, which we assume has V words. Then for each review document d, we draw a topic distribution $\theta^d \sim \text{Dir}(\alpha)$ as in standard LDA. For each sentence s in document d, we draw an aspect assignment $z_{d,s} \sim \text{Multi}(\theta^d)$.

Now for each word in sentence s of document d, we have several choices: The word may describe the specific aspect (e.g. *waiter* for the *staff* aspect), or a general aspect (e.g. *restaurant*), or an opinion either specific to the aspect (e.g. *friendly*) or generic (e.g. *great*), or a commonly used background word (e.g. *know*). To distinguish between these choices, we introduce two indicator variable, $y_{d,s,n}$ and $u_{d,s,n}$, for the nth word $w_{d,s,n}$. We draw $y_{d,s,n}$ from a multinomial distribution over $\{0, 1, 2\}$, parameterized by $\pi^{d,s,n}$. $y_{d,s,n}$ determines whether $w_{d,s,n}$ is a background word, aspect word or opinion word. We will discuss how to set $\pi^{d,s,n}$ in Section 3.2. We draw $u_{d,s,n}$ from a Bernoulli distribution over $\{0, 1\}$ parameterized by p, which in turn is drawn from a symmetric Beta(γ). $u_{d,s,n}$ determines whether $w_{d,s,n}$ is general or aspect-specific. We then draw $w_{d,s,n}$ as follows:

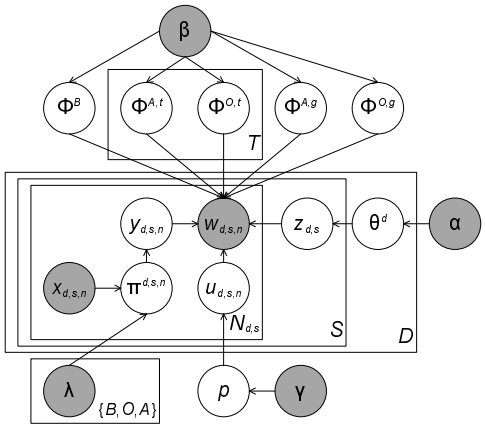

Figure 1: The plate notation of our model.

$$w_{d,s,n} \sim \begin{cases} \text{Multi}(\phi^{\mathcal{B}}) & \text{if } y_{d,s,n} = 0 \\ \text{Multi}(\phi^{\mathcal{A},z_{d,s}}) & \text{if } y_{d,s,n} = 1, u_{d,s,n} = 0 \\ \text{Multi}(\phi^{\mathcal{A},g}) & \text{if } y_{d,s,n} = 1, u_{d,s,n} = 1 \\ \text{Multi}(\phi^{\mathcal{O},z_{d,s}}) & \text{if } y_{d,s,n} = 2, u_{d,s,n} = 0 \\ \text{Multi}(\phi^{\mathcal{O},g}) & \text{if } y_{d,s,n} = 2, u_{d,s,n} = 1 \end{cases}.$$

Figure 1 shows our model using the plate notation.

3.2 Setting π with a Maximum Entropy Model

A simple way to set $\pi^{d,s,n}$ is to draw it from a symmetric Dirichlet prior. However, as suggested in (Mei et al., 2007; Lin and He, 2009), fully unsupervised topic models are unable to identify opinion words well. An important observation we make is that aspect words and opinion words usually play different syntactic roles in a sentence. Aspect words tend to be nouns while opinion words tend to be adjectives. Their contexts in sentences can also be different. But we do not want to use strict rules to separate aspect and opinion words because there are also exceptions. E.g. verbs such as *recommend* can also be opinion words.

In order to use information such as POS tags to help discriminate between aspect and opinion words, we propose a novel idea as follows: We set $\pi^{d,s,n}$ using a maximum entropy (MaxEnt) model applied to a feature vector $\boldsymbol{x}_{d,s,n}$ associated with $w_{d,s,n}$. $\boldsymbol{x}_{d,s,n}$ can encode any arbitrary features we think may be discriminative, e.g. previous, current and next POS tags. Formally, we have

$$p(y_{d,s,n} = l | \boldsymbol{x}_{d,s,n}) = \pi_l^{d,s,n} = \frac{\exp\left(\lambda_l \cdot \boldsymbol{x}_{d,s,n}\right)}{\sum_{l'=0}^{2} \exp\left(\lambda_{l'} \cdot \boldsymbol{x}_{d,s,n}\right)},$$

where $\{\lambda_l\}_{l=0}^{2}$ denote the MaxEnt model weights and can be learned from a set of training sentences with labeled background, aspect and opinion words. This MaxEnt-LDA hybrid model is partially inspired by (Mimno and McCallum, 2008).

As for the features included in $\boldsymbol{x}$, currently we use two types of simple features: (1) lexical features which include the previous, the current and the next words $\{\mathrm{w}_{i-1}, \mathrm{w}_i, \mathrm{w}_{i+1}\}$, and (2) POS tag features which include the previous, the current and the next POS tags $\{\mathrm{POS}_{i-1}, \mathrm{POS}_i, \mathrm{POS}_{i+1}\}$.

3.3 Inference

We use Gibbs sampling to perform model inference. Due to the space limit, we leave out the derivation details and only show the sampling formulas. Note that the MaxEnt component is trained first independently of the Gibbs sampling procedure, that is, in Gibbs sampling, we assume that the λ parameters are fixed.

We use $\boldsymbol{w}$ to denote all the words we observe in the collection, $\boldsymbol{x}$ to denote all the feature vectors for these words, and $\boldsymbol{y}$, $\boldsymbol{z}$ and $\boldsymbol{u}$ to denote all the hidden variables. First, given the assignment of all other hidden variables, to sample a value for $z_{d,s}$, we use the following formula:

$$P(z_{d,s}=t|\boldsymbol{z}_{\neg(d,s)},\boldsymbol{y},\boldsymbol{u},\boldsymbol{w},\boldsymbol{x}) \propto \frac{c^d_{(t)}+\alpha}{c^d_{(\cdot)}+T\alpha} \times\left(\frac{\Gamma\left(c^{\mathcal{A},t}_{(\cdot)}+V\beta\right)}{\Gamma\left(c^{\mathcal{A},t}_{(\cdot)}+n^{\mathcal{A},t}_{(\cdot)}+V\beta\right)}\cdot\prod_{v=1}^{V}\frac{\Gamma\left(c^{\mathcal{A},t}_{(v)}+n^{\mathcal{A},t}_{(v)}+\beta\right)}{\Gamma\left(c^{\mathcal{A},t}_{(v)}+\beta\right)}\right) \times\left(\frac{\Gamma\left(c^{\mathcal{O},t}_{(\cdot)}+V\beta\right)}{\Gamma\left(c^{\mathcal{O},t}_{(\cdot)}+n^{\mathcal{O},t}_{(\cdot)}+V\beta\right)}\cdot\prod_{v=1}^{V}\frac{\Gamma\left(c^{\mathcal{O},t}_{(v)}+n^{\mathcal{O},t}_{(v)}+\beta\right)}{\Gamma\left(c^{\mathcal{O},t}_{(v)}+\beta\right)}\right).$$

Here $c^d_{(t)}$ is the number of sentences assigned to aspect t in document d, and $c^d_{(\cdot)}$ is the number of sentences in document d. $c^{\mathcal{A},t}_{(v)}$ is the number of times word v is assigned as an *aspect* word to aspect t, and $c^{\mathcal{O},t}_{(v)}$ is the number of times word v is assigned as an *opinion* word to aspect t. $c^{\mathcal{A},t}_{(\cdot)}$ is the total number of times any word is assigned as an aspect word to aspect t, and $c^{\mathcal{O},t}_{(\cdot)}$ is the total number of times any word is assigned as an opinion word to aspect t. All these counts represented by a c variable exclude sentence s of document d. $n^{\mathcal{A},t}_{(v)}$ is the number of times word v is assigned as an aspect word to aspect t in sentence s of document d, and similarly, $n^{\mathcal{O},t}_{(v)}$ is the number of times word v is assigned as an opinion word to aspect t in sentence s of document d.

Then, to jointly sample values for $y_{d,s,n}$ and $u_{d,s,n}$, we have

$$\begin{aligned}
&P(y_{d,s,n}=0|\boldsymbol{z},\boldsymbol{y}_{\neg(d,s,n)},\boldsymbol{u}_{\neg(d,s,n)},\boldsymbol{w},\boldsymbol{x})\\
&\propto\frac{\exp(\lambda_0\cdot\boldsymbol{x}_{d,s,n})}{\sum_{l'}\exp(\lambda_{l'}\cdot\boldsymbol{x}_{d,s,n})}\cdot\frac{c^{\mathcal{B}}_{(w_{d,s,n})}+\beta}{c^{\mathcal{B}}_{(\cdot)}+V\beta},\\
&P(y_{d,s,n}=l,u_{d,s,n}=b|\boldsymbol{z},\boldsymbol{y}_{\neg(d,s,n)},\boldsymbol{u}_{\neg(d,s,n)},\boldsymbol{w},\boldsymbol{x})\\
&\propto\frac{\exp(\lambda_l\cdot\boldsymbol{x}_{d,s,n})}{\sum_{l'}\exp(\lambda_{l'}\cdot\boldsymbol{x}_{d,s,n})}\cdot g(w_{d,s,n},z_{d,s},l,b),
\end{aligned}$$

where the function $g(v,t,l,b)$ ($1\le v\le V, 1\le t\le T, l\in\{1,2\}, b\in\{0,1\}$) is defined as follows:

$$g(v,t,l,b)=\begin{cases}\frac{c^{\mathcal{A},t}_{(v)}+\beta}{c^{\mathcal{A},t}_{(\cdot)}+V\beta}\cdot\frac{c_{(0)}+\gamma}{c_{(\cdot)}+2\gamma} & \text{if } l=1,b=0\\ \frac{c^{\mathcal{O},t}_{(v)}+\beta}{c^{\mathcal{O},t}_{(\cdot)}+V\beta}\cdot\frac{c_{(0)}+\gamma}{c_{(\cdot)}+2\gamma} & \text{if } l=2,b=0\\ \frac{c^{\mathcal{A},g}_{(v)}+\beta}{c^{\mathcal{A},g}_{(\cdot)}+V\beta}\cdot\frac{c_{(1)}+\gamma}{c_{(\cdot)}+2\gamma} & \text{if } l=1,b=1\\ \frac{c^{\mathcal{O},g}_{(v)}+\beta}{c^{\mathcal{O},g}_{(\cdot)}+V\beta}\cdot\frac{c_{(1)}+\gamma}{c_{(\cdot)}+2\gamma} & \text{if } l=2,b=1.\end{cases}.$$

Here the various c variables denote various counts excluding the nth word in sentence s of document d. Due to space limit, we do not give full explanation here.

4 Experiment Setup

To evaluate our MaxEnt-LDA hybrid model for jointly modeling aspect and opinion words, we used a restaurant review data set previously used in (Ganu et al., 2009; Brody and Elhadad, 2010) and a hotel review data set previously used in (Baccianella et al., 2009). We removed stop words and used the Stanford POS Tagger[2] to tag the two data sets. Only reviews that have no more than 50 sentences were used. We also kept another version of the data which includes the stop words for the purpose of extracting the contextual features included in $\boldsymbol{x}$. Some details of the data sets are given in Table 1.

For our hybrid model, we ran 500 iterations of Gibbs sampling. Following (Griffiths and Steyvers, 2004), we fixed the Dirichlet priors as follows: $\alpha =$

[2] http://nlp.stanford.edu/software/tagger.shtml

data set	restaurant	hotel
#tokens	1,644,923	1,097,739
#docs	52,574	14,443

Table 1: Some statistics of the data sets.

data set	#sentences	#tokens
restaurant	46	634
cell phone	125	4414
DVD player	180	3024

Table 2: Some statistics of the labeled training data.

$50/T$, $\beta = 0.1$ and $\gamma = 0.5$. We also experimented with other settings of these priors and did not notice any major difference. For MaxEnt training, we tried three labeled data sets: one that was taken from the restaurant data set and manually annotated by us[3], and two from the annotated data set used in (Wu et al., 2009). Note that the latter two were used for testing domain adaptation in Section 6.3. Some details of the training sets are shown in Table 2.

In our preliminary experiments, we also tried two variations of our MaxEnt-LDA hybrid model. (1) The first is a fully unsupervised model where we used a uniform Dirichlet prior for π. We found that this unsupervised model could not separate aspect and opinion words well. (2) The second is a bootstrapping version of the MaxEnt-LDA model where we used the predicted values of y as pseudo labels and re-trained the MaxEnt model iteratively. We found that this bootstrapping procedure did not boost the overall performance much and even hurt the performance a little in some cases. Due to the space limit we do not report these experiments here.

5 Evaluation

In this section we report the evaluation of our model. We refer to our MaxEnt-LDA hybrid model as *ME-LDA*. We also implemented a local version of the standard LDA method where each sentence is treated as a document. This is the model used in (Brody and Elhadad, 2010) to identify aspects, and we refer to this model as *LocLDA*.

Food	Staff	Order Taking	Ambience
chocolate	service	wait	room
dessert	food	waiter	dining
cake	staff	wait	tables
cream	**excellent**	order	bar
ice	**friendly**	minutes	place
desserts	**attentive**	seated	decor
coffee	**extremely**	waitress	scene
tea	waiters	reservation	space
bread	**slow**	asked	area
cheese	**outstanding**	told	table

Table 4: Sample aspects of the restaurant domain using LocLDA. Note that the words in bold are opinion words which are mixed with aspect words.

5.1 Qualitative Evaluation

For each of the two data sets, we show four sample aspects identified by ME-LDA in Table 3 and Table 5. Because the hotel domain is somehow similar to the restaurant domain, we used the labeled training data from the restaurant domain also for the hotel data set. From the tables we can see that generally aspect words are quite coherent and meaningful, and opinion words correspond to aspects very well. For comparison, we also applied LocLDA to the restaurant data set and present the aspects in Table 4. We can see that ME-LDA and LocLDA give similar aspect words. The major difference between these two models is that ME-LDA can sperate aspect words and opinion words, which can be very useful. ME-LDA is also able to separate general opinion words from aspect-specific ones, giving more informative opinion expressions for each aspect.

5.2 Evaluation of Aspects Identification

We also quantitatively evaluated the quality of the automatically identified aspects. Ganu et al. (2009) provide a set of annotated sentences from the restaurant data set, in which each sentence has been assigned one or more labels from a gold standard label set $\mathcal{S}$ = {Staff, Food, Ambience, Price, Anecdote, Misc}. To evaluate the quality of our aspect identification, we chose from the gold standard labels three major aspects, namely *Staff*, *Food* and *Ambience*. We did not choose the other aspects because (1) *Price* is often mixed with other aspects such as *Food*, and (2) *Anecdote* and *Misc* do not show clear

[3] We randomly selected 46 sentences for manual annotation.

Food		Staff		Order Taking		Ambience		General
Aspect	**Opinion**	**Aspect**	**Opinion**	**Aspect**	**Opinion**	**Aspect**	**Opinion**	**Opinion**
chocolate	good	service	friendly	table	seated	room	small	good
dessert	best	staff	attentive	minutes	asked	dining	nice	well
cake	great	food	great	wait	told	tables	beautiful	nice
cream	delicious	wait	nice	waiter	waited	bar	romantic	great
ice	sweet	waiter	good	reservation	waiting	place	cozy	better
desserts	hot	place	excellent	order	long	decor	great	small
coffee	amazing	waiters	helpful	time	arrived	scene	open	bad
tea	fresh	restaurant	rude	hour	rude	space	warm	worth
bread	tasted	waitress	extremely	manager	sat	area	feel	definitely
cheese	excellent	waitstaff	slow	people	finally	table	comfortable	special

Table 3: Sample aspects and opinion words of the restaurant domain using ME-LDA.

Service		Room Condition		Ambience		Meal		General
Aspect	**Opinion**	**Aspect**	**Opinion**	**Aspect**	**Opinion**	**Aspect**	**Opinion**	**Opinion**
staff	helpful	room	shower	room	quiet	breakfast	good	great
desk	friendly	bathroom	small	floor	open	coffee	fresh	good
hotel	front	bed	clean	hotel	small	fruit	continental	nice
english	polite	air	comfortable	noise	noisy	buffet	included	well
reception	courteous	tv	hot	street	nice	eggs	hot	excellent
help	pleasant	conditioning	large	view	top	pastries	cold	best
service	asked	water	nice	night	lovely	cheese	nice	small
concierge	good	rooms	safe	breakfast	hear	room	great	lovely
room	excellent	beds	double	room	overlooking	tea	delicious	better
restaurant	rude	bath	well	terrace	beautiful	cereal	adequate	fine

Table 5: Sample aspects and opinion words of the hotel domain using ME-LDA.

patterns in either word usage or writing styles, making it even hard for humans to identify them. Brody and Elhadad (2010) also only used these three aspects for quantitative evaluation. To avoid ambiguity, we used only the single-labeled sentences for evaluation. About 83% of the labeled sentences have a single label, which confirms our observation that a sentence usually belongs to a single aspect.

We first ran ME-LDA and LocLDA each to get an inferred aspect set $\mathcal{T}$. Following (Brody and Elhadad, 2010), we set the number of aspects to 14 in both models. We then manually mapped each inferred aspect to one of the six gold standard aspects, i.e., we created a mapping function $f(t) : \mathcal{T} \rightarrow \mathcal{S}$. For sentence s of document d, we first assign it to an inferred aspect as follows:

$$t^* = \arg\max_{t \in \mathcal{T}} \sum_{n=1}^{N_{d,s}} \log P(w_{d,s,n}|t).$$

We then assign the gold standard aspect $f(t^*)$ to this sentence. We then calculated the F-1 score of the three aspects: *Staff*, *Food* and *Ambience*. The results are shown in Table 6. Generally ME-LDA has given competitive results compared with LocLDA. For *Food* and *Ambience* ME-LDA outperformed LocLDA, while for *Staff* ME-LDA is a little worse than LocLDA. Note that ME-LDA is not designed to compete with LocLDA for aspect identification.

Aspect	Method	Precision	Recall	F-1
Staff	LocLDA	0.804	0.585	0.677
	ME-LDA	0.779	0.540	0.638
Food	LocLDA	0.898	0.648	0.753
	ME-LDA	0.874	0.787	0.828
Ambience	LocLDA	0.603	0.677	0.638
	ME-LDA	0.773	0.558	0.648

Table 6: Results of aspects identification on restaurant.

5.3 Evaluation of Opinion Identification

Since the major advantage of ME-LDA is its ability to separate aspect and opinion words, we further quantitatively evaluated the quality of the aspect-specific opinion words identified by ME-LDA. Brody and Elhadad (2010) has constructed a gold standard set of aspect-specific opinion words for the restaurant data set. In this gold standard set, they manually judged eight out of the 14 automatically inferred aspects they had: $\mathcal{J}$ = {Ambiance, Staff, Food-Main Dishes, Atmosphere-Physical, Food-Baked Goods, Food-General, Drinks, Service}. Each word is assigned a polarity score ranging from -2.0 to 2.0 in each aspect. We used their gold standard words whose polarity scores are not equal to zero. Because their gold standard only includes adjectives, we also manually added more opinion words into the gold standard set. To do so, we took the top 20 opinion words returned by our method and two baseline methods, pooled them together, and manually judged them. We use precision at n (P@n), a commonly used metric in information retrieval, for evaluation. Because top words are more important in opinion models, we set n to 5, 10 and 20. For both ME-LDA and BL-1 below, we again manually mapped each automatically inferred aspect to one of the gold standard aspects.

Since LocLDA does not identify aspect-specific opinion words, we consider the following two baseline methods that can identify aspect-specific opinion words:

BL-1: In this baseline, we start with all adjectives as candidate opinion words, and use mutual information (MI) to rank these candidates. Specifically, given an aspect t, we rank the candidate words according to the following scoring function:

$$\text{Score}_{\text{BL-1}}(w,t) = \sum_{v \in \mathcal{V}_t} p(w,v) \log \frac{p(w,v)}{p(w)p(v)},$$

where $\mathcal{V}_t$ is the set of the top-100 frequent aspect words from $\phi^{A,t}$.

BL-2: In this baseline, we first use LocLDA to learn a topic distribution for each sentence. We then assign a sentence to the aspect with the largest probability and hence get sentence clusters. We manually map these clusters to the eight gold standard aspects. Finally, for each aspect we rank adjectives by their frequencies in the aspect and treat these as aspect-specific opinion words.

Method	P@5	P@10	P@20
ME-LDA	$0.825^{*,\diamond}$	0.700^{*}	0.569^{*}
BL-1	0.400	0.450	0.469
BL-2	0.725	0.650	0.563

Table 7: Average P@n of aspect-specific opinion words on restaurant. * and $\diamond$ indicate that the improvement hypothesis is accepted at confidence level 0.9 respectively for BL-1 and BL-2.

The basic results in terms of the average precision at n over the eight aspects are shown in Table 7. We can see that ME-LDA outperformed the two baselines consistently. Especially, for P@5, ME-LDA gave more than 100% relative improvement over BL-1. The absolute value of 0.825 for P@5 also indicates that top opinion words discovered by our model are indeed meaningful.

5.4 Evaluation of the Association between Opinion Words and Aspects

The evaluation in the previous section shows that our model returns good opinion words for each aspect. It does not, however, directly judge how *aspect-specific* those opinion words are. This is because the gold standard created by (Brody and Elhadad, 2010) also includes general opinion words. E.g. *friendly* and *good* may both be judged to be opinion words for the *staff* aspect, but the former is more specific than the latter. We suspect that BL-2 has comparable performance with ME-LDA for this reason. So we further evaluated the association between opinion words and aspects by directly looking at how easy it is to infer the corresponding aspect by only looking at an aspect-specific opinion word. We selected four aspects for evaluation: *Ambiance*, *Staff*, *Food-Main Dishes* and *Atmosphere-Physical* . We chose these four aspects because they are quite different from each other and thus manual judgments on these four aspects can be more objective. For each aspect, similar to the pooling strategy in IR, we pooled the top 20 opinion words identified by BL-1, BL-2 and ME-LDA. We then asked two human assessors to assign an association score to each of these words as follows: If the word is closely associated with an aspect, a score of 2 is given; if it is marginally as-

Metrics	Dataset	BL-2	ME-LDA
nDCG@5	Restaurant	0.647	0.764
	Hotel	0.782	0.820
nDCG@10	Restaurant	0.781	0.897
	Hotel	0.722	0.789

Table 8: Average nDCG performance of BL-2 and ME-LDA. Because only four aspects were used for evaluation, we did not perform statistical significance test. We found that in all cases ME-LDA outperformed BL-2 for either all aspects or three out of four aspects.

Methods	Average F-1
LocLDA	0.690
ME-LDA + $\mathcal{A}$	0.631
ME-LDA + $\mathcal{B}$	0.695
ME-LDA + $\mathcal{C}$	0.705

Table 9: Comparison of the average F-1 using different feature sets for aspect identification on restaurant.

sociated with an aspect, a score of 1 is given; otherwise, 0 is given. We calculated the Kappa statistics of agreement, and we got a quite high Kappa value of 0.8375 and 0.7875 respectively for the restaurant data set and the hotel data set. Then for each word in an aspect, we took the average of the scores of the two assessors. We used an nDCG-like metric to compare the performance of our model and of BL-2. The metric is defined as follows:

$$\text{nDCG@}k(t, \mathcal{M}) = \frac{\sum_{i=1}^{k} \frac{\text{Score}(\mathcal{M}_{t,i})}{\log_2(i+1)}}{\text{iDCG@}k(t)},$$

where $\mathcal{M}_{t,i}$ is the ith aspect-specific opinion word inferred by method $\mathcal{M}$ for aspect t, Score($\mathcal{M}_{t,i}$) is the association score of this word, and iDCG@$k(t)$ is the score of the ideal DCG measure at k for aspect t, that is, the maximum DCG score assuming an ideal ranking. We chose $k = 5$ and $k = 10$. The average nDCG over the four aspects are presented in Table 8. We can see that ME-LDA outperformed BL-2 quite a lot for the restaurant data set, which conforms to our hypothesis that ME-LDA generates aspect-specific opinion words of stronger association with aspects. For the hotel data set, ME-LDA outperformed a little. This may be due to the fact that we used the restaurant training data for the hotel data set.

6 Further Analysis of MaxEnt

In this section, we perform some further evaluation and analysis of the MaxEnt component in our model.

6.1 Feature Selection

Previous studies have shown that simple POS features and lexical features can be very effective for discovering aspect words and opinion words (Hu and Liu, 2004; Jin et al., 2009; Wu et al., 2009; Brody and Elhadad, 2010). for POS features, since we observe that aspect words tend to be nouns while opinion words tend to be adjectives but sometimes also verbs or other part-of-speeches, we can expect that POS features should be quite useful. As for lexical features, words from a sentiment lexicon can also be helpful in discovering opinion words.

However, lexical features are more diverse so presumably we need more training data in order to detect useful lexical features. Lexical features are also more domain-dependent. On the other hand, we hypothesize that POS features are more effective when the amount of training data is small and/or the training data comes from a different domain. We therefore compare the following three sets of features:

- $\mathcal{A}$: w_{i-1}, w_i, w_{i+1}
- $\mathcal{B}$: POS_{i-1}, POS_i, POS_{i+1}
- $\mathcal{C}$: $\mathcal{A} + \mathcal{B}$

We show the comparison of the performance in Table 9 using the average F-1 score defined in Section 5.2 for aspect identification, and in Table 10 using the average P@n measure defined in Section 5.3 for opinion identification. We can see that Set $\mathcal{B}$ plays the most important part, which conforms to our hypothesis that POS features are very important in opinion mining. In addition, we can see that Set $\mathcal{C}$ performs a bit better than Set $\mathcal{B}$, which indicates that some lexical features (e.g., general opinion words) may also be helpful. Note that here the training data is from the same domain as the test data, and therefore lexical features are likely to be useful.

6.2 Examine the Size of Labeled Data

As we have seen, POS features play the major role in discriminating between aspect and opinion words. Because there are much fewer POS features than word features, we expect that we do not need many

Methods	P@5	P@10	P@20
BL-2	0.725	0.650	0.563
ME-LDA + $\mathcal{A}$	0.150	0.200	0.231
ME-LDA + $\mathcal{B}$	0.775	0.688	0.569
ME-LDA + $\mathcal{C}$	0.825	0.700	0.569

Table 10: Comparison of the average P@n using different feature sets for opinion identification on restaurant.

Method	F-1
LocalLDA	0.690
ME-LDA + 10	0.629
ME-LDA + 20	0.692
ME-LDA + 30	0.691
ME-LDA + 40	0.726
ME-LDA + 46	0.705

Table 11: Average F-1 with differen sizes of training data on restaurant.

labeled sentences to learn the POS-based patterns. We now examine the sensitivity of the performance with respect to the amount of labeled data. We generated four smaller training data sets with 10, 20, 30 and 40 sentences each from the whole training data set we have, which consists of 46 labeled sentences. The results are shown in Table 11 and Table 12. We can see that generally the performance stays above BL when the number of training sentences is 20 or more. This indicates that our model needs only a relatively small number of high-quality training sentences to achieve good results.

6.3 Domain Adaption

Since we find that the MaxEnt supervision relies more on POS features than lexical features, we also hypothesize that if the training sentences come from a different domain the performance can still remain relatively high. To test this hypothesis, we tried two

Method	P@5	P@10	P@20
BL-2	0.725	0.650	0.563
ME-LDA + 10	0.700	0.563	0.488
ME-LDA + 20	0.875	0.650	0.600
ME-LDA + 30	0.825	0.700	0.569
ME-LDA + 40	0.825	0.688	0.581
ME-LDA + 46	0.825	0.700	0.569

Table 12: Average P@n of aspect-specific opinion words with differen sizes of training data on restaurant.

Method	Average F-1
restaurant + $\mathcal{B}$	0.695
restaurant + $\mathcal{C}$	0.705
cell phone + $\mathcal{B}$	0.662
cell phone + $\mathcal{C}$	0.629
DVD player + $\mathcal{B}$	0.686
DVD player + $\mathcal{C}$	0.635

Table 13: Average F-1 performance for domain adaption on restaurant.

Method	P@5	P@10	P@20
restaurant + $\mathcal{B}$	0.775	0.688	0.569
restaurant + $\mathcal{C}$	0.825	0.700	0.569
cell phone + $\mathcal{B}$	0.775	0.675	0.588
cell phone + $\mathcal{C}$	0.750	0.688	0.594
DVD player + $\mathcal{B}$	0.775	0.713	0.575
DVD player + $\mathcal{C}$	0.825	0.663	0.588

Table 14: Average P@n of aspect-specific opinion words for domain adaption on restaurant.

quite different training data sets, one from the *cell phone* domain and the other from the *DVD player* domain, both used in (Wu et al., 2009).

We consider two feature sets defined in Section 6.1 for domain adaption, namely $\mathcal{B}$ and $\mathcal{C}$. The results are shown in Table 13 and Table 14.

For aspect identification, using out-of-domain training data performed worse than using in-domain training data, but the absolute performance is still decent. And interestingly, we can see that using $\mathcal{B}$ is better than using $\mathcal{C}$, indicating that lexical features may hurt the performance in the cross-domain setting. It suggests that lexical features are not easily adaptable across domains for aspect identification.

For opinion identification, we can see that there is no clear difference between using out-of-domain training data and using in-domain training data, which may indicate that our opinion identification component is robust in domain adaption. Also, we cannot easily tell whether $\mathcal{B}$ has advantage over $\mathcal{C}$ for opinion identification. One possible reason may be that those general opinion words are useful across domains, so lexical features may still be useful for domain adaption.

7 Conclusions

In this paper, we presented a topic modeling approach that can jointly identify aspect and opinion words, using a MaxEnt-LDA hybrid. We showed that by incorporating a supervised, discriminative maximum entropy model into an unsupervised, generative topic model, we could leverage syntactic features to help separate aspect and opinion words. We evaluated our model on two large review data sets from the restaurant and the hotel domains. We found that our model was competitive in identifying meaningful aspects compared with previous models. Most importantly, our model was able to identify meaningful opinion words strongly associated with different aspects. We also demonstrated that the model could perform well with a relatively small amount of training data or with training data from a different domain.

Our model provides a principled way to jointly model both aspects and opinions. One of the future directions we plan to explore is to use this model to help sentence-level extraction of specific opinions and their targets, which previously was only tackled in a fully supervised manner. Another direction is to extend the model to support polarity classification.

ACKNOWLEDGMENT

The authors Xin Zhao, Hongfei Yan and Xiaoming Li are partially supported by NSFC under the grant No. 70903008 and 60933004, CNGI grant No. 2008-122, 863 Program No. 2009AA01Z143, and the Open Fund of the State Key Laboratory of Software Development Environment under Grant No. SKLSDE-2010KF-03, Beihang University.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet rating of product reviews. In *Proceedings of the 31st ECIR*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wei Jin and Hung Hay Ho. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. In *Proceedings of the 26th International Conference on Machine Learning*.

Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. OpinionMiner: A novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the Eighteenth ACM Conference on Information and Knowledge Management*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*.

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Conference on Uncertainty in Artificial Intelligence*.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the HLT-EMNLP*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th International Conference on World Wide Web*.

Yuanbin Wu, Qi Zhang, Xuangjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Jun Zhu and Eric P. Xing. 2010. Conditional topic random fields. In *Proceedings of the 27th International Conference on Machine Learning*.