Aspect Extraction through Semi-Supervised Modeling

Arjun Mukherjee
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
arjun4787@gmail.com

Bing Liu
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
liub@cs.uic.edu

Abstract

Aspect extraction is a central problem in sentiment analysis. Current methods either extract aspects without categorizing them, or extract and categorize them using unsupervised topic modeling. By categorizing, we mean the synonymous aspects should be clustered into the same category. In this paper, we solve the problem in a different setting where the user provides some seed words for a few aspect categories and the model extracts and clusters aspect terms into categories simultaneously. This setting is important because categorizing aspects is a subjective task. For different application purposes, different categorizations may be needed. Some form of user guidance is desired. In this paper, we propose two statistical models to solve this seeded problem, which aim to discover exactly what the user wants. Our experimental results show that the two proposed models are indeed able to perform the task effectively.

1 Introduction

Aspect-based sentiment analysis is one of the main frameworks for sentiment analysis (Hu and Liu, 2004; Pang and Lee, 2008; Liu, 2012). A key task of the framework is to extract aspects of entities that have been commented in opinion documents. The task consists of two sub-tasks. The first sub-task extracts *aspect terms* from an opinion corpus. The second sub-task clusters synonymous aspect terms into categories where each category represents a single aspect, which we call an *aspect category*. Existing research has proposed many methods for aspect extraction. They largely fall into two main types. The first type only extracts aspect terms without grouping them into categories (although a subsequent step may be used for the grouping, see Section 2). The second type uses statistical topic models to extract aspects and group them at the same time in an unsupervised manner. Both approaches are useful. However, in practice, one also encounters another setting, where grouping is not straightforward because for different applications the user may need different groupings to reflect the application needs. This problem was reported in (Zhai et al., 2010), which gave the following example. In car reviews, internal design and external design can be regarded as two separate aspects, but can also be regarded as one aspect, called "design", based on the level of details that the user wants to study. It is also possible that the same word may be put in different categories based on different needs. However, (Zhai et al., 2010) did not extract aspect terms. It only categorizes a set of given aspect terms.

In this work, we propose two novel statistical models to extract and categorize aspect terms automatically given some seeds in the user interested categories. It is thus able to best meet the user's specific needs. Our models also jointly model both aspects and aspect specific sentiments. The first model is called SAS and the second model is called ME-SAS. ME-SAS improves SAS by using Maximum-Entropy (or Max-Ent for short) priors to help separate aspects and sentiment terms. However, to train Max-Ent, we do not need manually labeled training data (see Section 4).

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 339–348,
Jeju, Republic of Korea, 8-14 July 2012.

In practical applications, asking users to provide some seeds is easy as they are normally experts in their trades and have a good knowledge what are important in their domains.

Our models are related to topic models in general (Blei et al., 2003) and joint models of aspects and sentiments in sentiment analysis in specific (e.g., Zhao et al., 2010). However, these current models are typically unsupervised. None of them can use seeds. With seeds, our models are thus semi-supervised and need a different formulation. Our models are also related to the DF-LDA model in (Andrzejewski et al., 2009), which allows the user to set must-link and cannot-link constraints. A must-link means that two terms must be in the same topic (aspect category), and a cannot-link means that two terms cannot be in the same topic. Seeds may be expressed with must-links and cannot-links constraints. However, our models are very different from DF-LDA. First of all, we jointly model aspect and sentiment, while DF-LDA is only for topics/aspects. Joint modeling ensures clear separation of aspects from sentiments producing better results. Second, our way of treating seeds is also different from DF-LDA. We discuss these and other related work in Section 2.

The proposed models are evaluated using a large number of hotel reviews. They are also compared with two state-of-the-art baselines. Experimental results show that the proposed models outperform the two baselines by large margins.

2 Related Work

There are many existing works on aspect extraction. One approach is to find frequent noun terms and possibly with the help of dependency relations (Hu and Liu, 2004; Popescu and Etzioni, 2005; Zhuang et al., 2006; Blair-Goldensohn et al., 2008; Ku et al., 2006; Wu et al., 2009; Somasundaran and Wiebe, 2009; Qiu et al., 2011). Another approach is to use supervised sequence labeling (Liu, Hu and Cheng 2005; Jin and Ho, 2009; Jakob and Gurevych, 2010; Li et al., 2010; Choi and Cardie, 2010; Kobayashi et al., 2007; Yu et al., 2011). Ma and Wan (2010) also exploited centering theory, and (Yi et al., 2003) used language models. However, all these methods do not group extracted aspect terms into categories. Although there are works on grouping aspect terms (Carenini et al., 2005; Zhai et al., 2010; Zhai et al., 2011; Guo et al., 2010), they all assume that aspect terms have been extracted beforehand.

In recent years, topic models have been used to perform extraction and grouping at the same time. Existing works are based on two basic models, pLSA (Hofmann, 1999) and LDA (Blei et al., 2003). Some existing works include discovering global and local aspects (Titov and McDonald, 2008), extracting key phrases (Branavan et al., 2008), rating multi-aspects (Wang et al., 2010; Moghaddam and Ester, 2011), summarizing aspects and sentiments (Lu et al., 2009), and modeling attitudes (Sauper et al., 2011). In (Lu and Zhai, 2008), a semi-supervised model was proposed. However, their method is entirely different from ours as they use expert reviews to guide the analysis of user reviews.

Aspect and sentiment extraction using topic modeling come in two flavors: discovering aspect words sentiment wise (i.e., discovering positive and negative aspect words and/or sentiments for each aspect without separating aspect and sentiment terms) (Lin and He, 2009; Brody and Elhadad, 2010, Jo and Oh, 2011) and separately discovering both aspects and sentiments (e.g., Mei et al., 2007; Zhao et al., 2010). Zhao et al. (2010) used Maximum-Entropy to train a switch variable to separate aspect and sentiment words. We adopt this method as well but with no use of manually labeled data in training. One problem with these existing models is that many discovered aspects are not understandable/meaningful to users. Chang et al. (2009) stated that one reason is that the objective function of topic models does not always correlate well with human judgments. Our seeded models are designed to overcome this problem.

Researchers have tried to generate "meaningful" and "specific" topics/aspects. Blei and McAuliffe (2007) and Ramage et al. (2009) used document label information in a supervised setting. Hu et al. (2011) relied on user feedback during Gibbs sampling iterations. Andrzejewski et al. (2011) incorporated first-order logic with Markov Logic Networks. However, it has a practical limitation for reasonably large corpora since the number of non-trivial groundings can grow to $O(N^2)$ where N is the number of unique tokens in the corpus. Andrzejewski et al. (2009) used another approach (DF-LDA) by introducing must-link and cannot-link constraints as Dirichlet Forest priors. Zhai et al. (2011) reported that the model does not scale up

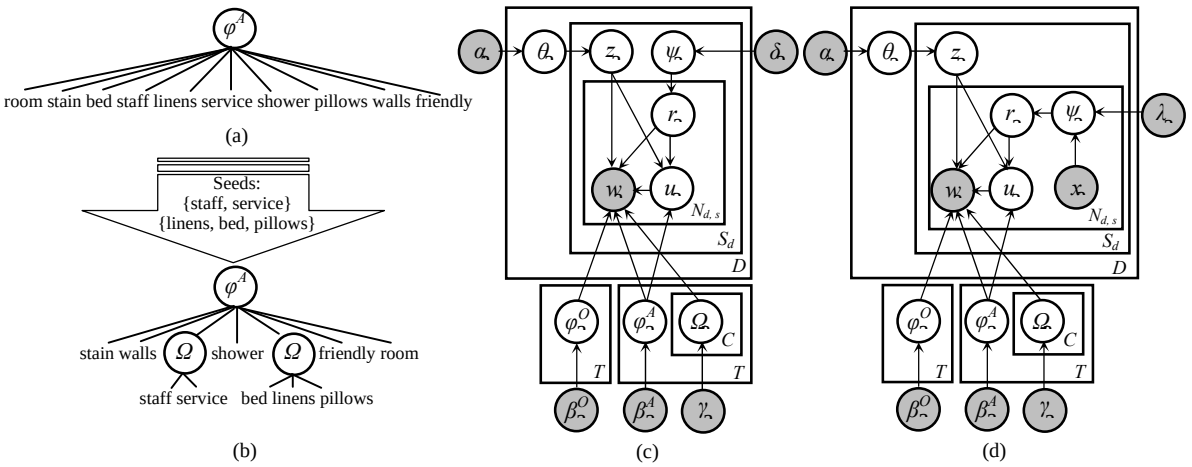

Figure 1: Prior structure: (a) Standard ASMs, (b) Two-level tree structured distribution. Graphical models in plate notation: (c) SAS and (d) ME-SAS.

when the number of cannot-links go beyond 1000 because the number of maximal cliques $Q^{(r)}$ in a connected component of size $|r|$ in the cannot-link graph is exponential in r. Note that we could still experiment with DF-LDA as our problem size is not so large. We will show in Section 4 that the proposed models outperform it by a large margin.

3 Proposed Seeded Models

The standard LDA and existing aspect and sentiment models (ASMs) are mostly governed by the phenomenon called "higher-order co-occurrence" (Heinrich, 2009), i.e., based on how often terms co-occur in different contexts[1]. This unfortunately results in many "non-specific" terms being pulled and clustered. We employ seed sets to address this issue by "guiding" the model to group semantically related terms in the same aspect thus making the aspect more specific and related to the seeds (which reflect the user needs). For easy presentation, we will use *aspect* to mean *aspect category* from now on. We replace the multinomial distribution over words for each aspect (as in ASMs) with a special two-level tree structured distribution. The generative process of ASMs assumes that each vocabulary word is independently (i.e., not dependent upon other word-aspect association) and equally probable to be associated with any aspect. Due to higher-order co-occurrences, we find conceptually different terms yet related in contexts (e.g., in hotel domain terms like stain, shower, walls in aspect *Maintenance*; bed, linens, pillows in aspect *Cleanliness*) equally probable of emission for any aspect. Figure 1(a) shows an example tree. Upon adding the seed sets {bed, linens, pillows} and {staff, service}, the prior structure now changes to the correlated distribution in Figure 1 (b). Thus, each aspect has a top level distribution over non-seed words and seed sets. Each seed set in each aspect further has a second level distribution over seeds in that seed set. The aspect term (word) emission now requires two steps: first sampling at level one to obtain a non-seed word or a seed set. If a non-seed word is sampled we emit it else we further sample at the second seed set level and emit a seed word. This ensures that seed words together have either all high or low aspect associations. Furthermore, seed sets preserve conjugacy between related concepts and also shape more specific aspects by clustering based on higher order co-occurrences with seeds rather than only with standard one level multinomial distribution over words (or terms) alone.

3.1 SAS Model

We now present the proposed Seeded Aspect and Sentiment model (SAS). Let $v_{1\ldots V}$ denote the entries in our vocabulary where V is the number of unique non-seed terms. Let there be C seed sets $Q_{l=1\ldots C}$ where each seed set Q_l is a group of semantically related terms. Let $\varphi^A_{t=1\ldots T}, \varphi^O_{t=1\ldots T}$ denote T aspect and aspect specific sentiment models. Also let $\Omega_{t,l}$ denote the aspect specific distribution of seeds in the seed set Q_l. Following the approach of (Zhao et al., 2010), we too assume that a review sentence usually talks about one

[1] w_1 co-occurring with w_2 which in turn co-occurs with w_3 denotes a second-order co-occurrence between w_1 and w_3.

aspect. A review document $d_{1\ldots D}$ comprises of S_d sentences and each sentence $s \in S_d$ has $N_{d,s}$words. Also, let $Sent_s^d$ denote the sentence s of document d. To distinguish between aspect and sentiment terms, we introduce an indicator (switch) variable $r_{d,s,j} \in \{\hat{a}, \hat{o}\}$ for the j^{th}term of $Sent_s^d, w_{d,s,j.}$ Further, let $\psi_{d,s}$ denote the distribution of aspects and sentiments in $Sent_s^d$. The generative process of the SAS model (see Figure 1(c)) is given by:

1. For each aspect $t \in \{1, \ldots, T\}$:
   i. Draw $\varphi_t^O \sim Dir(\beta^O)$
   ii. Draw a distribution over terms and seed sets $\varphi_t^A \sim Dir(\beta^A)$
      a) For each seed set $l \in \{Q_1, \ldots, Q_C\}$
         Draw a distribution over seeds $\Omega_{t,l} \sim Dir(\gamma)$
2. For each (review) document $d \in \{1, \ldots, D\}$:
   i. Draw $\theta_d \sim Dir(\alpha)$
   ii. For each sentence $s \in \{1, \ldots, S_d\}$:
      a) Draw $z_{d,s} \sim Mult(\theta_d)$
      b) Draw $\psi_{d,s} \sim Beta(\delta)$
      c) For each term $w_{d,s,j}$ where $j \in \{1, \ldots, N_{d,s}\}$:
         I. Draw $r_{d,s,j} \sim Bernoulli(\psi_{d,s})$, $r_{d,s,j} \in \{\hat{a}, \hat{o}\}$
         II. if $r_{d,s,j} = \hat{o}$ // $w_{d,s,j}$ is a sentiment
            Emit $w_{d,s,j} \sim Mult(\varphi^O_{z_{d,s}})$
            else // $r_{d,s,j} = \hat{a}$, $w_{d,s,j}$ is an aspect
            A. Draw $u_{d,s,j} \sim Mult(\varphi^A_{z_{d,s}})$
            B. if $u_{d,s,j} \in V$ // non-seed term
               Emit $w_{d,s,j} = u_{d,s,j}$
               else // $u_{d,s,j}$ is some seed set index say $l_{d,s,j}$
               Emit $w_{d,s,j} \sim \Omega_{z_{d,s},l_{d,s,j}}$

We employ collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for posterior inference. As z and r are at different hierarchical levels, we derive their samplers separately as follows:

$$p(z_{d,s} = t \,|Z_{\neg d,s}, R_{\neg d,s}, W_{\neg d,s}, U_{\neg d,s}) \propto \frac{B(n^O_{t,[]} + \beta^O)}{B\left(n^O_{t,[]_{\neg d,s}} + \beta^O\right)} \times \frac{B(n^{U,A}_{t,[]} + \beta^A)}{B(n^{U,A}_{t,[]_{\neg d,s}} + \beta^A)} \times \prod_{l=1}^{C} \frac{B(n^{S,A}_{t,l,[]} + \gamma)}{B(n^{S,A}_{t,l,[]_{\neg d,s}} + \gamma)} \times \frac{n^{Sent.}_{d,t\ \ \neg d,s} + \alpha}{n^{Sent.}_{d,(\cdot)\ \ \neg d,s} + T\alpha} \quad (1)$$

$$p(r_{d,s,j} = \hat{o} | Z_{\neg d,s}, R_{\neg d,s,j}, W_{\neg d,s,j}, U_{\neg d,s,j}, z_{d,s} = t, w_{d,s,j} = w) \propto \frac{n^O_{t,w\ \neg d,s,j} + \beta^O}{n^O_{t,(\cdot)\ \neg d,s,j} + |V \cup \cup_l Q_l|\beta^O} \times \frac{n^O_{d,s\ \neg d,s,j} + \delta_b}{n^A_{d,s\ \neg d,s,j} + \delta_a + n^O_{d,s\ \neg d,s,j} + \delta_b} \quad (2)$$

$$p(r_{d,s,j} = \hat{a} | \ldots) \propto \begin{cases} \frac{n^{S,A}_{t,l,w\ \neg d,s,j} + \gamma}{n^{S,A}_{t,l,(\cdot)\ \neg d,s,j} + |Q_l|\gamma} \times \frac{n^{\Omega}_{t,l} + \beta^A}{n^{\Omega}_{t,(\cdot)} + (V+C)\beta^A} \times \frac{n^A_{d,s\ \neg d,s,j} + \delta_b}{n^A_{d,s\ \neg d,s,j} + \delta_a + n^O_{d,s\ \neg d,s,j} + \delta_b} & ;\ w \in Q_l \\ \frac{n^{U,A}_{t,w} + \beta^A}{n^{U,A}_{t,(\cdot)} + (V+C)\,\beta^A} \times \frac{n^A_{d,s\ \neg d,s,j} + \delta_b}{n^A_{d,s\ \neg d,s,j} + \delta_a + n^O_{d,s\ \neg d,s,j} + \delta_b} & ; \nexists l, w \in Q_l \end{cases} \quad (3)$$

where $B(\vec{x}) = \frac{\prod_{i=1}^{\dim(\vec{x})} \Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\vec{x})} x_i)}$is the multinomial Beta function. $n^O_{t,v}$ is the number of times term v was assigned to aspect t as an opinion/sentiment word. $n^{U,A}_{t,v}$ is the number of times non-seed term $v \in V$ was assigned to aspect t as an aspect. $n^{S,A}_{t,l,v}$ is the number of times seed term $v \in V_l$ was assigned to aspect t as an aspect. $n^{Sent.}_{d,t}$ is the number of sentences in document d that were assigned to aspect t. $n^A_{d,s}$ and $n^O_{d,s}$ denote the number of terms in $Sent_s^d$ that were assigned to aspects and opinions respectively. $n^{\Omega}_{t,l}$ is the number of times any term of seed set Q_l was assigned to aspect t. Omission of a latter index denoted by [] in the above notation represents the corresponding row vector spanning over the latter index. For example, $n^{U,A}_{t,[]} = [n^{U,A}_{t,v=1}, \ldots, n^{U,A}_{t,v=V}]$ and $(\cdot)$ denotes the marginalized sum over the latter index. The subscript $\neg d, s$ denotes the counts excluding assignments of all terms in $Sent_s^d$. $\neg d, s, j$ denotes counts excluding $w_{d,s,j.}$We perform hierarchical sampling. First, an aspect is sampled for each sentence $z_{d,s}$ using Eq. (1). After sampling the aspect, we sample $r_{d,s,j}$. The probability of $w_{d,s,j}$ being an opinion or sentiment term, $p(r_{d,s,j} = \hat{o})$ is given by Eq. (2). However, for $p(r_{d,s,j} = \hat{a})$ we have two cases: (a) the observed term $w = w_{d,s,j} \in Q_l$ or (b) does not belong to any seed set, $\nexists l, w \in Q_l$, i.e., w is an non-seed term. These cases are dealt in Eq. (3).

Asymmetric Beta priors: Hyper-parameters α, β^O, β^A are not very sensitive and the heuristic values suggested in (Griffiths and Steyvers, 2004) usually hold well in practice (Wallach et al. 2009). However, the smoothing hyper-parameter δ (Figure 1(c)) is crucial as it governs the aspect or sentiment switch. Essentially, $\psi_{d,s} \sim Beta(\delta\vec{\xi})$ is the probability of emitting an aspect term[2] in $Sent_s^d$ with concentration parameter δ and base measure $\vec{\xi} = [\xi_a, \xi_b]$. Without any prior belief, uniform base measures $\xi_a = \xi_b = 0.5$ are used resulting in symmetric Beta priors. However, aspects are often more probable than sentiments in a sentence (e.g., "*The beds, sheets, and bedding were dirty*."). Thus, it is more principled to employ asymmetric priors. Using a labeled set of sentences, $S_{labeled}$, where we know the per sentence probability of aspect emission ($\psi_{d,s}$), we can employ the method of moments to estimate the smoothing hyper-parameter $\delta = [\delta_a, \delta_b]$:

$$\delta_a = \mu\left(\frac{\mu(1-\mu)}{\sigma} - 1\right), \delta_b = \delta_a\left(\frac{1}{\mu} - 1\right);\ \mu = E[\psi_{d,s}], \sigma = Var[\psi_{d,s}] \quad (4)$$

[2] $r_{d,s,j} \sim Bernoulli(\psi_{d,s})$. $\psi_{d,s}$, $1 - \psi_{d,s}$ are the success and failure probability of emitting an aspect/sentiment term.

3.2 ME-SAS Model

We can further improve SAS by employing Maximum Entropy (Max-Ent) priors for aspect and sentiment switching. We call this new model ME-SAS. The motivation is that aspect and sentiment terms play different syntactic roles in a sentence. Aspect terms tend to be nouns or noun phrases while sentiment terms tend to be adjectives, adverbs, etc. POS tag information can be elegantly encoded by moving $\psi_{d,s}$ to the term plate (see Figure 1(d)) and drawing it from a Max-Ent($x_{d,s,j}; \lambda$) model. Let $\overrightarrow{x_{d,s,j}} = [POS_{w_{d,s,j-1}}, POS_{w_{d,s,j}}, POS_{w_{d,s,j+1}}, w_{d,s,j} - 1, w_{d,s,j}, w_{d,s,j} + 1]$ denote the feature vector associated with $w_{d,s,j}$ encoding lexical and POS features of the previous, current and next term. Using a training data set, we can learn Max-Ent priors. Note that unlike traditional Max-Ent training, we do not need manually labeled data for training (see Section 4 for details). For ME-SAS, only the sampler for the switch variable *r* changes as follows:

$$p(r_{d,s,j} = \hat{o} | Z_{\neg d,s}, R_{\neg d,s,j}, W_{\neg d,s,j}, U_{\neg d,s,j}, z_{d,s} = t, w_{d,s,j} = w) \propto \frac{n^{O}_{t,w_{\neg d,s,j}} + \beta^{O}}{n^{O}_{t,(\cdot)_{\neg d,s,j}} + |V \cup \cup_l Q_l| \beta^{O}} \times \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, \hat{o}))}{\sum_{y \in \{\hat{a}, \hat{o}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, y))} \quad (5)$$

$$p(r_{d,s,j} = \hat{a} | \ldots) \propto \begin{cases} \frac{n^{S,A}_{t,l,w_{\neg d,s,j}} + \gamma}{n^{S,A}_{t,l,(\cdot)_{\neg d,s,j}} + |Q_l| \gamma} \times \frac{n^{\Omega}_{t,l} + \beta^{A}}{n^{\Omega}_{t,(\cdot)} + (V+C)\beta^{A}} \times \\ \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, \hat{a}))}{\sum_{y \in \{\hat{a}, \hat{o}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, y))} \quad ; \; w \in Q_l \\ \frac{n^{U,A}_{t,w} + \beta^{A}}{n^{U,A}_{t,(\cdot)} + (V+C)\beta^{A}} \times \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, \hat{a}))}{\sum_{y \in \{\hat{a}, \hat{o}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, y))} ; \nexists l, w \in Q_l \end{cases} \quad (6)$$

where $\lambda_{1\ldots n}$ are the parameters of the learned Max-Ent model corresponding to the n binary feature functions $f_{1\ldots n}$ of Max-Ent.

4 Experiments

This section evaluates the proposed models. Since the focus in this paper is to generate high quality aspects using seeds, we will not evaluate sentiments although both SAS and ME-SAS can also discover sentiments. To compare the performance with our models, we use two existing state-of-the-art models, ME-LDA (Zhao et al. 2010) and DF-LDA (Andrzejewski et al., 2009). As discussed in Section 2, there are two main flavors of aspect and sentiment models. The first flavor does not separate aspect and sentiment, and the second flavor uses a switch to perform the separation. Since our models also perform a switch, it is natural to compare with the latter flavor, which is also more advanced. ME-LDA is the representative model in this flavor. DF-LDA adds constraints to LDA. We use our seeds to generate constraints for DF-LDA. While ME-LDA cannot consider constraints, DF-LDA does not separate sentiments and aspects. Apart from other modeling differences, our models can do both, which enable them to produce much better results.

Dataset and Settings: We used hotel reviews from tripadvisor.com. Our corpus consisted of 101,234 reviews and 692,783 sentences. Punctuations, stop words [3], and words appearing less than 5 times in the corpus were removed.

For all models, the posterior inference was drawn after 5000 Gibbs iterations with an initial burn-in of 1000 iterations. For SAS and ME-SAS, we set $\alpha = 50/T$, $\beta^A = \beta^O = 0.1$ as suggested in (Griffiths and Steyvers, 2004). To make the seeds more effective, we set the seed set word-distribution hyper-parameter γ to be much larger than β^A, the hyper-parameter for the distribution over seed sets and aspect terms. This results in higher weights to seeded words which in turn guide the sampler to cluster relevant terms better. A more theoretical approach would involve performing hyper-parameter estimation (Wallach et al., 2009) which may reveal specific properties of the dataset like the estimate of α (indicating how different documents are in terms of their latent semantics), β (suggesting how large the groups of frequently appearing aspect and sentiment terms are) and γ (giving a sense of which and how large groupings of seeds are good). These are interesting questions and we defer it to our future work. In this work, we found that the setting $\gamma = 250$, a larger value compared to β^A, produced good results.

For SAS, the asymmetric Beta priors were estimated using the method of moments (Section 3.1). We sampled 500 random sentences from the corpus and for each sentence identified the aspects. We thus computed the per-sentence probability of aspect emission ($\psi_{d,s}$) and used Eq. (4) to compute the final estimates, which give $\delta_a = 2.35$, $\delta_b = 3.44$.

To learn the Max-Ent parameters λ of ME-SAS, we used the sentiment lexicon [4] of (Hu and Liu, 2004) to automatically generate training data (no manual labeling). We randomly sampled 1000 terms from the corpus which have appeared at least

[3] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[4] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

Aspect (seeds)	ME-SAS		SAS		ME-LDA		DF-LDA
	Aspect	Sentiment	Aspect	Sentiment	Aspect	Sentiment	Topic
Staff (staff service waiter hospitality upkeep)	attendant manager waitress maintenance bartender waiters housekeeping receptionist waitstaff janitor	friendly attentive polite nice **clean** pleasant **slow** courteous rude professional	attendant waiter waitress manager maintenance **helpful** waiters housekeeping receptionist **polite**	friendly nice dirty **comfortable** nice **clean** polite **extremely** courteous efficient	staff maintenance **room** upkeep **linens** room-service receptionist **wait** **pillow** waiters	friendly nice courteous **extremely** nice **clean** polite **little** helpful **better**	staff ***friendly*** ***helpful*** **beds** **front** room **comfortable** **large** receptionist housekeeping
Cleanliness (curtains restroom floor beds cleanliness)	carpets hall towels bathtub couch mattress linens wardrobe spa pillow	clean dirty comfortable fresh wet filthy extra stain **front** worn	hall carpets towels pillow stain mattress **filthy** linens **interior** bathtub	clean dirty fresh old nice good **enough** new **front** **friendly**	cleanliness floor carpets bed **lobby** bathroom **staff** closet spa **décor**	clean good dirty **hot** **large** nice **fresh** **thin** new **little**	clean pool **beach** carpets **parking** bed bathroom **nice** comfortable suite
Comfort (comfort mattress furniture couch pillows)	bedding bedcover sofa linens bedroom suites **décor** comforter blanket futon	comfortable clean soft nice uncomfortable spacious hard comfy dirty quiet	bed linens sofa bedcover **hard** bedroom **privacy** **double** **comfy** futon	nice dirty comfortable large clean **best** spacious **only** big **extra**	bed mattress suites furniture **lighting** **décor** room bedroom **hallway** **carpet**	great clean awesome dirty best comfortable soft nice **only** **extra**	bed mattress **nice** ***stay*** **lighting** lobby comfort room **dirty** sofa

Table 1: Top ranked aspect and sentiment words in three aspects (please see the explanation in Section 4.1).

20 times (to ensure that the training set is reasonably representative of the corpus). Of those 1000 terms if they appeared in the sentiment lexicon, they were treated as sentiment terms, else aspect terms. Clearly, labeling words not in the sentiment lexicon as aspect terms may not always be correct. Even with this noisy automatically-labeled data, the proposed models can produce good results. Since ME-LDA used manually labeled training data for Max-Ent, we again randomly sampled 1000 terms from our corpus appearing at least 20 times and labeled them as aspect terms or sentiment terms, so this labeled data clearly has less noise than our automatically labeled data. For both ME-SAS and ME-LDA we used the corresponding feature vector of each labeled term (in the context of sentences where it occurs) to train the Max-Ent model. As DF-LDA requires must-link and cannot-link constraints, we used our seed sets to generate intra-seed set must-link and inter-seed set cannot-link constraints. For its hyper-parameters, we used the default values in the package[5] (Andrzejewski et al., 2009).

Setting the number of topics/aspects in topic models is often tricky as it is difficult to know the exact number of topics that a corpus has. While non-parametric Bayesian approaches (Teh et al., 2006) do exist for estimating the number of topics, T, they strongly depend on the hyper-parameters (Heinrich, 2009). As we use fixed hyper-parameters, we do not learn T from Bayesian non-parametrics. We used 9 major aspects (T = 9) based on commonsense knowledge of what people usually talk about hotels and some experiments. These are *Dining*, *Staff*, *Maintenance*, *Check In*, *Cleanliness*, *Comfort*, *Amenities*, *Location* and *Value for Money* (VFM). However, it is important to note that the proposed models are flexible and do not need to have seeds for every aspect/topic. Our experiments simulate the real-life situation where the user may not know all aspects or have no seeds for some aspects. Thus, we provided seeds only to the first 6 of the 9 aspects/topics. We will see that without seeds for all aspects, our models not only can improve the seeded aspects but also improve the non-seeded aspects.

4.1 Qualitative Results

This section shows some qualitative results to give an intuitive feeling of the results from different models. Table 1 shows the aspect terms and sentiment terms discovered by the 4 models for

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

three aspects. Due to space limitations, we are unable to show all 6 aspects for which we have seeds. Since DF-LDA cannot separate aspects and sentiments, we only show its topics (aspects). **Red** (**bold**) colored words show semantic clustering errors or inappropriate terms for different groups.

It is important to note that we judge the results based on how they are related to the user seeds (which represent the user need). The judgment is to some extent subjective. What we reported here are based on our judgments what are appropriate and what are not for each aspect. For SAS, ME-SAS and ME-LDA, we mark sentiment terms as errors when they are grouped under aspects as these models are supposed to separate sentiments and aspects. For DF-LDA, the situation is different as it is not meant to separate sentiment and aspect terms, we use ***red*** italic font to indicate those adjectives which are aspect specific adjectives (see more discussion below). Our judgment may be slightly unfair to ME-LDA and DF-LDA as their results may make sense in some other ways. However, that is precisely the purpose of this work, to produce results that suit the user's need rather than something generic.

We can see from Table 1 that ME-SAS performs the best. Next in order are SAS, ME-LDA, and DF-LDA. We see that only providing a handful of seeds (5) for the aspect *Staff*, ME-SAS can discover highly specific words like manager, attendant, bartender, and janitor. By specific, we mean they are highly related to the given seeds. While SAS also discovers specific words benefiting from seeds, relying on Beta priors for aspect and sentiment switching was less effective. Next in performance is ME-LDA which although produces reasonable results in general, several aspect terms are far from what the user wants based on the seeds, e.g., room, linens, wait, pillow. Finally, we observe that DF-LDA does not perform well either. One reason is that it is unable to separate aspects and sentiments. Although encoding the intra-seed set must-link and inter-seed set cannot-link constraints in DF-LDA discovers some specific words as ME-SAS, they are much lower in the ranked order and hence do not show up in the top 10 words in Table 1. As DF-LDA is not meant to perform extraction and to group both aspect and sentiment terms, we relax the errors of DF-LDA due to correct aspect specific sentiments (e.g., friendly, helpful for *Staff* are correct aspect specific sentiments, but still regard incorrect sentiments like front, comfortable, large as errors) placed in aspect models. We call this model DF-LDA-Relaxed.

4.2 Quantitative Results

Topic models are often evaluated quantitatively using perplexity and likelihood on held-out test data (Blei et al., 2003). However, perplexity does not reflect our purpose since our aim is not to predict whether an unseen document is likely to be a review of some particular aspect. Nor are we trying to evaluate how well the unseen review data fits our seeded models. Instead our focus is to evaluate how well our learned aspects perform in clustering specific terms guided by seeds. So we directly evaluate the discovered aspect terms. Note again we do not evaluate sentiment terms as they are not the focus of this paper [6]. Since aspects produced by the models are rankings and we do not know the number of correct aspect terms, a natural way to evaluate these rankings is to use *precision @ n* (or *p@n*), where *n* is a rank position.

Varying number of seeds: Instead of a fixed number of seeds, we want to see the effect of the number of seeds on aspect discovery. Table 2 reports the average *p@n* vs. the number of seeds. The average is a two-way averaging. The first average was taken over all combinations of actual seeds selected for each aspect, e.g., when the number of seeds is 3, out of the 5 seeds in each aspect, all $\binom{5}{3}$ combinations of seeds were tried and the results averaged. The results were further averaged over *p@n* for 6 aspects with seeds. We start with 2 seeds and progressively increase them to 5. Using only 1 seed per seed set (or per aspect) has practically no effect because the top level distribution φ^A encodes which seed sets (and non-seed words) to include; the lower-level distribution Ω constrains the probabilities of the seed words to be correlated for each of the seed sets. Thus, having only one seed per seed set will result in sampling that single word whenever that seed set is chosen which will not have the effect of correlating seed words so as to pull other words based on co-occurrence with constrained seed words. From Table 2, we can see that for all models *p@n* progressively improves as the number of seeds increases. Again ME-SAS performs the best followed by SAS and DF-LDA.

[6] A qualitative evaluation of sentiment extraction based on Table 1 yields the following order: ME-SAS, SAS, ME-LDA.

No. of Seeds	DF-LDA			DF-LDA-Relaxed			SAS			ME-SAS		
	P@10	P@20	P@30	P@10	P@20	P@30	P@10	P@20	P@30	P@10	P@20	P@30
2	0.51	0.53	0.49	0.67	0.69	0.70	0.69	0.71	0.67	0.74	0.72	0.70
3	0.53	0.54	0.50	0.71	0.70	0.71	0.71	0.72	0.70	0.78	0.75	0.72
4	0.57	0.56	0.53	0.73	0.73	0.73	0.75	0.74	0.73	0.83	0.79	0.76
5	0.59	0.57	0.54	0.75	0.74	0.75	0.77	0.76	0.74	0.86	0.81	0.77

Table 2: Average *p@n* of the seeded aspects with the no. of seeds.

Aspect	ME-LDA			DF-LDA			DF-LDA-Relaxed			SAS			ME-SAS		
	P@10	P@20	P@30	P@10	P@20	P@30	P@10	P@20	P@30	P@10	P@20	P@30	P@10	P@20	P@30
Dining	0.70	0.65	0.67	0.50	0.60	0.63	0.70	0.70	0.70	0.80	0.75	0.73	0.90	0.85	0.80
Staff	0.60	0.70	0.67	0.40	0.65	0.60	0.60	0.75	0.67	0.80	0.80	0.70	1.00	0.90	0.77
Maintenance	0.80	0.75	0.73	0.40	0.55	0.56	0.60	0.70	0.73	0.70	0.75	0.76	0.90	0.85	0.80
Check In	0.70	0.70	0.67	0.50	0.65	0.60	0.80	0.75	0.70	0.80	0.70	0.73	0.90	0.80	0.76
Cleanliness	0.70	0.75	0.67	0.70	0.70	0.63	0.70	0.75	0.70	0.80	0.75	0.70	1.00	0.85	0.83
Comfort	0.60	0.70	0.63	0.60	0.65	0.50	0.70	0.75	0.63	0.60	0.75	0.67	0.90	0.80	0.73
Amenities	0.80	0.80	0.67	0.70	0.65	0.53	0.90	0.75	0.73	0.90	0.80	0.70	1.00	0.85	0.73
Location	0.60	0.70	0.63	0.50	0.60	0.56	0.70	0.70	0.67	0.60	0.70	0.63	0.70	0.75	0.67
VFM	0.50	0.55	0.50	0.40	0.50	0.46	0.60	0.60	0.60	0.50	0.50	0.50	0.60	0.55	0.53
Avg.	0.67	0.70	0.65	0.52	0.62	0.56	0.70	0.72	0.68	0.72	0.72	0.68	0.88	0.80	0.74

Table 3: Effect of performance on seeded and non-seeded aspects (5 seeds were used for the 6 seeded aspects).

Effect of seeds on non-seeded aspects: Here we compare all models aspect wise and see the results of seeded models SAS and ME-SAS on non-seeded aspects (Table 3). Shaded cells in Table 3 give the *p@n* values for DF-LDA, DF-LDA-Relaxed, SAS, and ME-SAS on three non-seeded aspects (Amenities, Location, and VFM)[7].

We see that across all the first 6 aspects with (5) seeds ME-SAS outperforms all other models by large margins in all top 3 ranked buckets *p*@10, *p*@20 and *p*@30. Next in order are SAS, ME-LDA and DF-LDA. For the last three aspects which did not have any seed guidance, we find something interesting. Seeded models SAS and especially ME-SAS result in improvements of non-seeded aspects too. This is because as seeds facilitate clustering specific and appropriate terms in seeded aspects, which in turn improves precision on non-seeded aspects. This phenomenon can be clearly seen in Table 1. In aspect *Staff* of ME-LDA, we find *pillow* and *linens* being clustered. This is not a "flaw" of the model per se, but the point here is *pillow* and *linens* happen to co-occur many times with other words like *maintenance*, *staff*, and *upkeep* because "room-service" generally includes staff members coming and replacing linens and pillow covers. Although *pillow* and *linens* are related to *Staff*, strictly speaking they are semantically incorrect because they do not represent the very concept "Staff" based on the seeds (which reflect the user need). Presence of seed sets in SAS and ME-SAS result in pulling such words as linens and pillow (due to seeds like beds and cleanliness in the aspect *Cleanliness*) and ranking them higher in the aspect *Cleanliness* (see Table 1) where they make more sense than *Staff*. Lastly, we also note that the improvements in non-seeded aspects are more pronounced for ME-SAS than SAS as SAS encounters more switching errors which counters the improvement gained by seeds.

In summary, the averages over all aspects (Table 3 last row) show that the proposed seeded models SAS and ME-SAS outperform ME-LDA, DF-LDA and even DF-LDA-Relaxed considerably.

5 Conclusion

This paper studied the issue of using seeds to discover aspects in an opinion corpus. To our knowledge, no existing work deals with this problem. Yet, it is important because in practice the user often has something in mind to find. The results obtained in a completely unsupervised manner may not suit the user's need. To solve this problem, we proposed two models SAS and ME-SAS which take seeds reflecting the user needs to discover specific aspects. ME-SAS also does not need any additional help from the user in its Max-Ent training. Our results showed that both models outperformed two state-of-the-art existing models ME-LDA and DF-LDA by large margins.

Acknowledgments

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

[7] Note that Tables 2 and 3 are different runs of the model. The variations in the results are due to the random initialization of the Gibbs sampler.

References

Andrzejewski, D., Zhu, X. and Craven, M. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. Proceedings of International Conference on Machine Learning (ICML).

Andrzejewski, D., Zhu, X. and Craven, M. and Recht, B. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. Proceedings of the 22nd International Joint Conferences on Artificial Intelligence (IJCAI).

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A. and Reynar, J. 2008. Building a sentiment summarizer for local service reviews. Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era.

Blei, D., Ng, A. and Jordan, M. 2003. Latent dirichlet allocation. The Journal of Machine Learning Research 3: 993-1022.

Blei D. and McAuliffe, J. 2007. Supervised topic models. Neural Information Processing Systems (NIPS).

Branavan, S., Chen, H., Eisenstein J. and Barzilay, R. 2008. Learning document-level semantic properties from free-text annotations. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

Brody, S. and Elhadad, S. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. Proceedings of The 2010 Annual Conference of the North American Chapter of the ACL (NAACL).

Carenini, G., Ng, R. and Zwart, E. 2005. Extracting knowledge from evaluative text. Proceedings of Third Intl. Conf. on Knowledge Capture (K-CAP-05).

Chang, J., Boyd-Graber, J., Wang, C. Gerrish, S. and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. In Neural Information Processing Systems (NIPS).

Choi, Y. and Cardie, C. 2010. Hierarchical sequential learning for extracting opinions and their attributes. Proceedings of Annual Meeting of the Association for Computational (ACL).

Griffiths, T. and Steyvers, M. 2004. Finding scientific topics. Proceedings of National Academy of Sciences (PNAS).

Guo, H., Zhu, H., Guo, Z., Zhang, X. and Su, X. 2009. Product feature categorization with multilevel latent semantic association. Proceedings of ACM International Conference on Information and Knowledge Management (CIKM).

Heinrich, G. 2009. A Generic Approach to Topic Models. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD).

Hofmann, T. 1999. Probabilistic latent semantic indexing. Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI).

Hu, Y., Boyd-Graber, J. and Satinoff, B. 2011. Interactive topic modeling. Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2011.

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. International Conference on Knowledge Discovery and Data Mining (ICDM).

Jakob, N. and Gurevych, I. 2010. Extracting Opinion Targets in a Single-and Cross-Domain Setting with Conditional Random Fields. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).

Jin, W. and Ho, H. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. Proceedings of International Conference on Machine Learning (ICML).

Jo, Y. and Oh, A. 2011. Aspect and sentiment unification model for online review analysis. ACM Conference in Web Search and Data Mining (WSDM).

Kobayashi, N., Inui, K. and Matsumoto, K. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Ku, L., Liang, Y. and Chen, H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. Proceedings of AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW'06).

Li, F., Han, C., Huang, M., Zhu, X. Xia, Y., Zhang, S. and Yu, H. 2010. Structure-aware review mining and summarization. International Conference on Computational Linguistics (COLING).

Lin, C. and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. Proceedings of ACM International Conference on Information and Knowledge Management (CIKM).

Liu, B. 2012. Sentiment Analysis and Opinion Mining.

Morgan & Claypool publishers (to appear in June 2012).

Liu, B, M. Hu, and J. Cheng. 2005. Opinion Observer: Analyzing and comparing opinions on the web. Proceedings of International Conference on World Wide Web (WWW).

Lu, Y., Zhai, C. and Sundaresan, N. 2009. Rated aspect summarization of short comments. Proceedings of International Conference on World Wide Web (WWW).

Lu, Y. and Zhai, C. 2008. Opinion Integration Through Semi-supervised Topic Modeling. Proceedings of the 17th International World Wide Web Conference (WWW).

Ma, T. and Wan, X. 2010. Opinion target extraction in Chinese news comments. Proceedings of Coling 2010 Poster Volume (COLING).

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. Proceedings of International Conference on World Wide Web (WWW).

Moghaddam, S. and Ester, M. 2011. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. Proceedings of the Annual ACM SIGIR International conference on Research and Development in Information Retrieval (SIGIR).

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval.

Popescu, A. and Etzioni, O. 2005. Extracting product features and opinions from reviews. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).

Qiu, G., Liu, B., Bu, J. and Chen, C. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Sauper, C., Haghighi, A. and Barzilay, R. 2011. Content models with attitude. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL).

Somasundaran, S. and Wiebe, J. 2009. Recognizing stances in online debates, Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP.

Teh, Y., Jordan, M., Beal, M. and Blei, D. 2006. Hierarchical Dirichlet Processes. In Journal of the American Statistical Association (JASA).

Titov, I. and McDonald, R. 2008. Modeling online reviews with multi-grain topic models. Proceedings of International Conference on World Wide Web (WWW).

Wallach, H., Mimno, D. and McCallum, A. 2009. Rethinking LDA: Why priors matter. In Neural Information Processing Systems (NIPS).

Wang, H., Lu, Y. and Zhai, C. 2010. Latent aspect rating analysis on review text data: a rating regression approach. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Wu, Y., Zhang, Q., Huang, X. and Wu, L. 2009. Phrase dependency parsing for opinion mining. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).

Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. Proceedings of IEEE International Conference on Data Mining (ICDM).

Yu, J., Zha, Z. J., Wang, M. and Chua, T. S. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL).

Zhai, Z., Liu, B. Xu, H. and Jia, P. 2010. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints. Proceedings of International Conference on Computational Linguistics (COLING).

Zhai, Z., Liu, B. Xu, H. and Jia, P. 2011. Constrained LDA for Grouping Product Features in Opinion Mining. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD).

Zhao, W., Jiang, J., Yan, Y. and Li, X. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).

Zhuang, L., Jing, F. and Zhu, X. 2006. Movie review mining and summarization. Proceedings of International Conference on Information and Knowledge Management (CIKM).