A Joint Model of Text and Aspect Ratings for Sentiment Summarization

Ivan Titov
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
titov@uiuc.edu

Ryan McDonald
Google Inc.
76 Ninth Avenue
New York, NY 10011
ryanmcd@google.com

Abstract

Online reviews are often accompanied with numerical ratings provided by users for a set of service or product aspects. We propose a statistical model which is able to discover corresponding topics in text and extract textual evidence from reviews supporting each of these aspect ratings – a fundamental problem in aspect-based sentiment summarization (Hu and Liu, 2004a). Our model achieves high accuracy, without any explicitly labeled data except the user provided opinion ratings. The proposed approach is general and can be used for segmentation in other applications where sequential data is accompanied with correlated signals.

1 Introduction

User generated content represents a unique source of information in which user interface tools have facilitated the creation of an abundance of labeled content, e.g., topics in blogs, numerical product and service ratings in user reviews, and helpfulness rankings in online discussion forums. Many previous studies on user generated content have attempted to predict these labels automatically from the associated text. However, these labels are often present in the data already, which opens another interesting line of research: designing models leveraging these labelings to improve a wide variety of applications.

In this study, we look at the problem of *aspect-based sentiment summarization* (Hu and Liu, 2004a; Popescu and Etzioni, 2005; Gamon et al., 2005; Carenini et al., 2006; Zhuang et al., 2006).[1] An aspect-based summarization system takes as input a set of user reviews for a specific product or service and produces a set of relevant aspects, the aggregated sentiment for each aspect, and supporting textual evidence. For example, figure 1 summarizes a restaurant using aspects *food*, *decor*, *service*, and *value* plus a numeric rating out of 5.

Nikos' Fine Dining

Food	4/5	"Best fish in the city", "Excellent appetizers"
Decor	3/5	"Cozy with an old world feel", "Too dark"
Service	1/5	"Our waitress was rude", "Awful service"
Value	5/5	"Good Greek food for the $", "Great price!"

Figure 1: An example aspect-based summary.

Standard aspect-based summarization consists of two problems. The first is *aspect identification and mention extraction*. Here the goal is to find the set of relevant aspects for a rated entity and extract all textual mentions that are associated with each. Aspects can be fine-grained, e.g., *fish*, *lamb*, *calamari*, or coarse-grained, e.g., *food*, *decor*, *service*. Similarly, extracted text can range from a single word to phrases and sentences. The second problem is *sentiment classification*. Once all the relevant aspects and associated pieces of texts are extracted, the system should aggregate sentiment over each aspect to provide the user with an average numeric or symbolic rating. Sentiment classification is a well studied problem (Wiebe, 2000; Pang et al., 2002; Turney, 2002) and in many domains users explicitly

[1] We use the term *aspect* to denote properties of an object that can be rated by a user as in Snyder and Barzilay (2007). Other studies use the term *feature* (Hu and Liu, 2004b).

Proceedings of ACL-08: HLT, pages 308–316,
Columbus, Ohio, USA, June 2008.

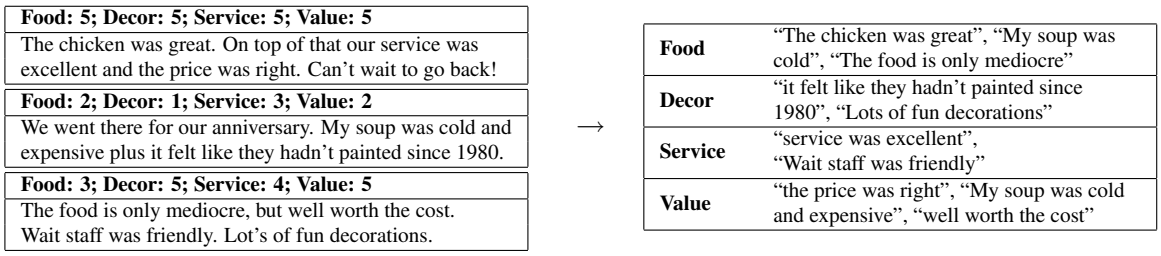


Figure 2: Extraction problem: Produce aspect mentions from a corpus of aspect rated reviews.

provide ratings for each aspect making automated means unnecessary.[2] Aspect identification has also been thoroughly studied (Hu and Liu, 2004b; Gamon et al., 2005; Titov and McDonald, 2008), but again, ontologies and users often provide this information negating the need for automation.

Though it may be reasonable to expect a user to provide a rating for each aspect, it is unlikely that a user will annotate every sentence and phrase in a review as being relevant to some aspect. Thus, it can be argued that the most pressing challenge in an aspect-based summarization system is to extract all relevant mentions for each aspect, as illustrated in figure 2. When labeled data exists, this problem can be solved effectively using a wide variety of methods available for text classification and information extraction (Manning and Schutze, 1999). However, labeled data is often hard to come by, especially when one considers all possible domains of products and services. Instead, we propose an unsupervised model that leverages aspect ratings that frequently accompany an online review.

In order to construct such model, we make two assumptions. First, ratable aspects normally represent coherent topics which can be potentially discovered from co-occurrence information in the text. Second, we hypothesize that the most predictive features of an aspect rating are features derived from the text segments discussing the corresponding aspect. Motivated by these observations, we construct a joint statistical model of text and sentiment ratings. The model is at heart a topic model in that it assigns words to a set of induced *topics*, each of which may represent one particular aspect. The model is extended through a set of maximum entropy classifiers, one per each rated aspect, that are used to predict the sentiment rating towards each of the aspects. However, only the words assigned to an aspects corresponding topic are used in predicting the rating for that aspect. As a result, the model enforces that words assigned to an aspects' topic are predictive of the associated rating. Our approach is more general than the particular statistical model we consider in this paper. For example, other topic models can be used as a part of our model and the proposed class of models can be employed in other tasks beyond sentiment summarization, e.g., segmentation of blogs on the basis of topic labels provided by users, or topic discovery on the basis of tags given by users on social bookmarking sites.[3]

The rest of the paper is structured as follows. Section 2 begins with a discussion of the joint text-sentiment model approach. In Section 3 we provide both a qualitative and quantitative evaluation of the proposed method. We conclude in Section 4 with an examination of related work.

2 The Model

In this section we describe a new statistical model called the Multi-Aspect Sentiment model (MAS), which consists of two parts. The first part is based on Multi-Grain Latent Dirichlet Allocation (Titov and McDonald, 2008), which has been previously shown to build topics that are representative of ratable aspects. The second part is a set of sentiment predictors per aspect that are designed to force specific topics in the model to be directly correlated with a particular aspect.

2.1 Multi-Grain LDA

The Multi-Grain Latent Dirichlet Allocation model (MG-LDA) is an extension of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). As was demon-

[2] E.g., http://zagat.com and http://tripadvisor.com.

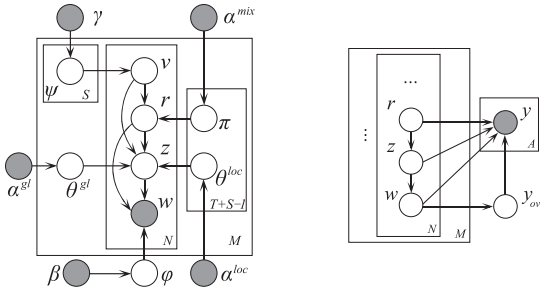
[3] See e.g. del.ico.us (http://del.ico.us).

strated in Titov and McDonald (2008), the topics produced by LDA do not correspond to ratable aspects of entities. In particular, these models tend to build topics that globally classify terms into product instances (e.g., Creative Labs Mp3 players versus iPods, or New York versus Paris Hotels). To combat this, MG-LDA models two distinct types of topics: global topics and local topics. As in LDA, the distribution of global topics is fixed for a document (a user review). However, the distribution of local topics is allowed to vary across the document.

A word in the document is sampled either from the mixture of global topics or from the mixture of local topics specific to the local context of the word. It was demonstrated in Titov and McDonald (2008) that ratable aspects will be captured by local topics and global topics will capture properties of reviewed items. For example, consider an extract from a review of a London hotel: "...public transport in London is straightforward, the tube station is about an 8 minute walk ...or you can get a bus for £1.50". It can be viewed as a mixture of topic *London* shared by the entire review (words: "London", "tube", "£"), and the ratable aspect *location*, specific for the local context of the sentence (words: "transport", "walk", "bus"). Local topics are reused between very different types of items, whereas global topics correspond only to particular types of items.

In MG-LDA a document is represented as a set of sliding windows, each covering T adjacent sentences within a document.[4] Each window v in document d has an associated distribution over local topics $\theta^{loc}_{d,v}$ and a distribution defining preference for local topics versus global topics $\pi_{d,v}$. A word can be sampled using any window covering its sentence s, where the window is chosen according to a categorical distribution $\psi_{d,s}$. Importantly, the fact that windows overlap permits the model to exploit a larger co-occurrence domain. These simple techniques are capable of modeling local topics without more expensive modeling of topic transitions used in (Griffiths et al., 2004; Wang and McCallum, 2005; Wallach, 2006; Gruber et al., 2007). Introduction of a symmetrical Dirichlet prior $Dir(\gamma)$ for the distribution $\psi_{d,s}$ can control the smoothness of transitions.

[4] Our particular implementation is over sentences, but sliding windows in theory can be over any sized fragment of text.



(a) (b)

Figure 3: (a) MG-LDA model. (b) An extension of MG-LDA to obtain MAS.

The formal definition of the model with K^{gl} global and K^{loc} local topics is as follows: First, draw K^{gl} word distributions for global topics φ^{gl}_z from a Dirichlet prior $Dir(\beta^{gl})$ and K^{loc} word distributions for local topics $\varphi^{loc}_{z'}$ - from $Dir(\beta^{loc})$. Then, for each document d:

- Choose a distribution of global topics $\theta^{gl}_d \sim Dir(\alpha^{gl})$.
- For each sentence s choose a distribution over sliding windows $\psi_{d,s}(v) \sim Dir(\gamma)$.
- For each sliding window v
  - choose $\theta^{loc}_{d,v} \sim Dir(\alpha^{loc})$,
  - choose $\pi_{d,v} \sim Beta(\alpha^{mix})$.
- For each word i in sentence s of document d
  - choose window $v_{d,i} \sim \psi_{d,s}$,
  - choose $r_{d,i} \sim \pi_{d,v_{d,i}}$,
  - if $r_{d,i} = gl$ choose global topic $z_{d,i} \sim \theta^{gl}_d$,
  - if $r_{d,i}{=}loc$ choose local topic $z_{d,i} {\sim} \theta^{loc}_{d,v_{d,i}}$,
  - choose word $w_{d,i}$ from the word distribution $\varphi^{r_{d,i}}_{z_{d,i}}$.

$Beta(\alpha^{mix})$ is a prior Beta distribution for choosing between local and global topics. In Figure 3a the corresponding graphical model is presented.

2.2 Multi-Aspect Sentiment Model

MG-LDA constructs a set of topics that ideally correspond to ratable aspects of an entity (often in a many-to-one relationship of topics to aspects). A major shortcoming of this model – and all other unsupervised models – is that this correspondence is not explicit, i.e., how does one say that topic X is really about aspect Y? However, we can observe that numeric aspect ratings are often included in our data by users who left the reviews. We then make the assumption that the text of the review discussing an aspect is predictive of its rating. Thus, if we model the prediction of aspect ratings jointly with the construction of explicitly associated topics, then such a

model should benefit from both higher quality topics and a direct assignment from topics to aspects. This is the basic idea behind the Multi-Aspect Sentiment model (MAS).

In its simplest form, MAS introduces a classifier for each aspect, which is used to predict its rating. Each classifier is explicitly associated to a single topic in the model and only words assigned to that topic can participate in the prediction of the sentiment rating for the aspect. However, it has been observed that ratings for different aspects can be correlated (Snyder and Barzilay, 2007), e.g., very negative opinion about room cleanliness is likely to result not only in a low rating for the aspect *rooms*, but also is very predictive of low ratings for the aspects *service* and *dining*. This complicates discovery of the corresponding topics, as in many reviews the most predictive features for an aspect rating might correspond to another aspect. Another problem with this overly simplistic model is the presence of opinions about an item in general without referring to any particular aspect. For example, "this product is the worst I have ever purchased" is a good predictor of low ratings for every aspect. In such cases, non-aspect 'background' words will appear to be the most predictive. Therefore, the use of the aspect sentiment classifiers based only on the words assigned to the corresponding topics is problematic. Such a model will not be able to discover coherent topics associated with each aspect, because in many cases the most predictive fragments for each aspect rating will not be the ones where this aspect is discussed.

Our proposal is to estimate the distribution of possible values of an aspect rating on the basis of the *overall* sentiment rating and to use the words assigned to the corresponding topic to compute *corrections* for this aspect. An aspect rating is typically correlated to the overall sentiment rating[5] and the fragments discussing this particular aspect will help to correct the overall sentiment in the appropriate direction. For example, if a review of a hotel is generally positive, but it includes a sentence "the neighborhood is somewhat seedy" then this sentence is predictive of rating for an aspect *location* being below other ratings. This rectifies the aforementioned problems. First, aspect sentiment ratings can often be regarded as conditionally independent given the overall rating, therefore the model will not be forced to include in an aspect topic any words from other aspect topics. Secondly, the fragments discussing overall opinion will influence the aspect rating only through the overall sentiment rating. The overall sentiment is almost always present in the real data along with the aspect ratings, but it can be coarsely discretized and we preferred to use a latent overall sentiment.

The MAS model is presented in Figure 3b. Note that for simplicity we decided to omit in the figure the components of the MG-LDA model other than variables r, z and w, though they are present in the statistical model. MAS also allows for extra unassociated local topics in order to capture aspects not explicitly rated by the user. As in MG-LDA, MAS has global topics which are expected to capture topics corresponding to particular types of items, such *London hotels* or *seaside resorts* for the hotel domain. In figure 3b we shaded the aspect ratings y_a, assuming that every aspect rating is present in the data (though in practice they might be available only for some reviews). In this model the distribution of the overall sentiment rating y_{ov} is based on all the n-gram features of a review text. Then the distribution of y_a, for every rated aspect a, can be computed from the distribution of y_{ov} and from any n-gram feature where at least one word in the n-gram is assigned to the associated aspect topic ($r = loc,\ z = a$).

Instead of having a latent variable y_{ov},[6] we use a similar model which does not have an explicit notion of y_{ov}. The distribution of a sentiment rating y_a for each rated aspect a is computed from two scores. The first score is computed on the basis of all the n-grams, but using a common set of weights independent of the aspect a. Another score is computed only using n-grams associated with the related topic, but an aspect-specific set of weights is used in this computation. More formally, we consider the log-linear distribution:

$$P(y_a = y|\mathbf{w}, \mathbf{r}, \mathbf{z}) \propto \exp(b^a_y + \sum_{f \in \mathbf{w}} J_{f,y} + p^a_{f,\mathbf{r},\mathbf{z}} J^a_{f,y}), \quad (1)$$

where **w**, **r**, **z** are vectors of all the words in a docu-

[5] In the dataset used in our experiments all three aspect ratings are equivalent for 5,250 reviews out of 10,000.

[6] Preliminary experiments suggested that this is also a feasible approach, but somewhat more computationally expensive.

ment, assignments of context (global or local) and topics for all the words in the document, respectively. b_y^a is the bias term which regulates the prior distribution $P(y_a = y)$, f iterates through all the n-grams, $J_{y,f}$ and $J_{y,f}^a$ are common weights and aspect-specific weights for n-gram feature f. $p_{f,\mathbf{r},\mathbf{z}}^a$ is equal to a fraction of words in n-gram feature f assigned to the aspect topic ($r = loc, z = a$).

2.3 Inference in MAS

Exact inference in the MAS model is intractable. Following Titov and McDonald (2008) we use a collapsed Gibbs sampling algorithm that was derived for the MG-LDA model based on the Gibbs sampling method proposed for LDA in (Griffiths and Steyvers, 2004). Gibbs sampling is an example of a Markov Chain Monte Carlo algorithm (Geman and Geman, 1984). It is used to produce a sample from a joint distribution when only conditional distributions of each variable can be efficiently computed. In Gibbs sampling, variables are sequentially sampled from their distributions conditioned on all other variables in the model. Such a chain of model states converges to a sample from the joint distribution. A naive application of this technique to LDA would imply that both assignments of topics to words $\mathbf{z}$ and distributions θ and φ should be sampled. However, (Griffiths and Steyvers, 2004) demonstrated that an efficient collapsed Gibbs sampler can be constructed, where only assignments $\mathbf{z}$ need to be sampled, whereas the dependency on distributions θ and φ can be integrated out analytically.

In the case of MAS we also use maximum a-posteriori estimates of the sentiment predictor parameters b_y^a, $J_{y,f}$ and $J_{y,f}^a$. The MAP estimates for parameters b_y^a, $J_{y,f}$ and $J_{y,f}^a$ are obtained by using stochastic gradient ascent. The direction of the gradient is computed simultaneously with running a chain by generating several assignments at each step and averaging over the corresponding gradient estimates. For details on computing gradients for log-linear graphical models with Gibbs sampling we refer the reader to (Neal, 1992).

Space constraints do not allow us to present either the derivation or a detailed description of the sampling algorithm. However, note that the conditional distribution used in sampling decomposes into two parts:

$$P(v_{d,i} = v, r_{d,i} = r, z_{d,i} = z|\mathbf{v'}, \mathbf{r'}, \mathbf{z'}, \mathbf{w}, \mathbf{y}) \propto \eta_{v,r,z}^{d,i} \times \rho_{r,z}^{d,i}, \quad (2)$$

where **v'**, **r'** and **z'** are vectors of assignments of sliding windows, context (global or local) and topics for all the words in the collection except for the considered word at position i in document d; y is the vector of sentiment ratings. The first factor $\eta_{v,r,z}^{d,i}$ is responsible for modeling co-occurrences on the window and document level and coherence of the topics. This factor is proportional to the conditional distribution used in the Gibbs sampler of the MG-LDA model (Titov and McDonald, 2008). The last factor quantifies the influence of the assignment of the word (d, i) on the probability of the sentiment ratings. It appears only if ratings are known (observable) and equals:

$$\rho_{r,z}^{d,i} = \prod_a \frac{P(y_a^d|\mathbf{w}, \mathbf{r'}, r_{d,i} = r, \mathbf{z'}, z_{d,i} = z)}{P(y_a^d|\mathbf{w}, \mathbf{r'}, \mathbf{z'}, r_{d,i} = gl)},$$

where the probability distribution is computed as defined in expression (1), y_a^d is the rating for the ath aspect of review d.

3 Experiments

In this section we present qualitative and quantitative experiments. For the qualitative analysis we show that topics inferred by the MAS model correspond directly to the associated aspects. For the quantitative analysis we show that the MAS model induces a distribution over the rated aspects which can be used to accurately predict whether a text fragment is relevant to an aspect or not.

3.1 Qualitative Evaluation

To perform qualitative experiments we used a set of reviews of hotels taken from TripAdvisor.com[7] that contained 10,000 reviews (109,024 sentences, 2,145,313 words in total). Every review was rated with at least three aspects: *service*, *location* and *rooms*. Each rating is an integer from 1 to 5. The dataset was tokenized and sentence split automatically.

[7] (c) 2005-06, TripAdvisor, LLC All rights reserved

	rated aspect	top words
local topics	service	staff friendly helpful service desk concierge excellent extremely hotel great reception english pleasant help
	location	hotel walk location station metro walking away right minutes close bus city located just easy restaurants
	rooms	room bathroom shower bed tv small water clean comfortable towels bath nice large pillows space beds tub
	-	breakfast free coffee internet morning access buffet day wine nice lobby complimentary included good fruit
	-	$ night parking rate price paid day euros got cost pay hotel worth euro expensive car extra deal booked
	-	room noise night street air did door floor rooms open noisy window windows hear outside problem quiet sleep
global topics	-	moscow st russian petersburg nevsky russia palace hermitage kremlin prospect river prospekt kempinski
	-	paris tower french eiffel dame notre rue st louvre rer champs opera elysee george parisian du pantheon cafes

Table 1: Top words from MAS for hotel reviews.

K_{rooms}	top words
2	rooms clean hotel room small nice comfortable modern good quite large lobby old decor spacious decorated bathroom size
	room noise night street did air rooms door open noisy window floor hear windows problem outside quiet sleep bit light
3	room clean bed comfortable rooms bathroom small beds nice large size tv spacious good double big space huge king
	room floor view rooms suite got views given quiet building small balcony upgraded nice high booked asked overlooking
	room bathroom shower air water did like hot small towels door old window toilet conditioning open bath dirty wall tub
4	room clean rooms comfortable bed small beds nice bathroom size large modern spacious good double big quiet decorated
	check arrived time day airport early room luggage took late morning got long flight ready minutes did taxi bags went
	room noise night street did air rooms noisy open door hear windows window outside quiet sleep problem floor conditioning
	bathroom room shower tv bed small water towels bath tub large nice toilet clean space toiletries flat wall sink screen

Table 2: Top words for aspect *rooms* with different number of topics K_{rooms}.

We ran the sampling chain for 700 iterations to produce a sample. Distributions of words in each topic were estimated as the proportion of words assigned to each topic, taking into account topic model priors β^{gl} and β^{loc}. The sliding windows were chosen to cover 3 sentences for all the experiments. All the priors were chosen to be equal to 0.1. We used 15 local topics and 30 global topics. In the model, the first three local topics were associated to the rating classifiers for each aspects. As a result, we would expect these topics to correspond to the *service*, *location*, and *rooms* aspects respectively. Unigram and bigram features were used in the sentiment predictors in the MAS model. Before applying the topic models we removed punctuation and also removed stop words using the standard list of stop words,[8] however, all the words and punctuation were used in the sentiment predictors.

It does not take many chain iterations to discover initial topics. This happens considerably faster than the appropriate weights of the sentiment predictor being learned. This poses a problem, because, in the beginning, the sentiment predictors are not accurate enough to force the model to discover appropriate topics associated with each of the rated aspects. And as soon as topic are formed, aspect sentiment predictors cannot affect them anymore because they do not have access to the true words associated with their aspects. To combat this problem we first train the sentiment classifiers by assuming that $p^a_{f,\mathbf{r},\mathbf{z}}$ is equal for all the local topics, which effectively ignores the topic model. Then we use the estimated parameters within the topic model.[9] Secondly, we modify the sampling algorithm. The conditional probability used in sampling, expression (2), is proportional to the product of two factors. The first factor, $\eta^{d,i}_{v,r,z}$, expresses a preference for topics likely from the co-occurrence information, whereas the second one, $\rho^{d,i}_{r,z}$, favors the choice of topics which are predictive of the observable sentiment ratings. We used $(\rho^{d,i}_{r,z})^{1+0.95^t q}$ in the sampling distribution instead of $\rho^{d,i}_{r,z}$, where t is the iteration number. q was chosen to be 4, though the quality of the topics seemed to be indistinguishable with any q between 3 and 10. This can be thought of as having $1 + 0.95^t q$ ratings instead of a single vector assigned to each review, i.e., focusing the model on prediction of the ratings rather than finding the topic labels which are good at explaining co-occurrences of words. These heuristics influence sampling only during the first iterations of the chain.

Top words for some of discovered local topics, in-

[8] http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

[9] Initial experiments suggested that instead of doing this 'pre-training' we could start with very large priors α^{loc} and α^{mix}, and then reduce them through the course of training. However, this is significantly more computationally expensive.

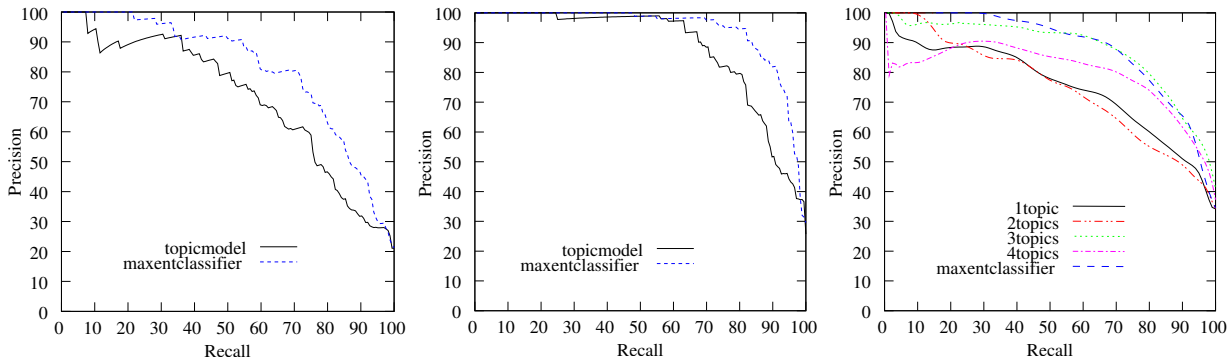


Figure 4: (a) Aspect *service*. (b) Aspect *location*. (c) Aspect *rooms*.

cluding the first 3 topics associated with the rated aspects, and also top words for some of global topics are presented in Table 1. We can see that the model discovered as its first three topics the correct associated aspects: *service*, *location*, and *rooms*. Other local topics, as for the MG-LDA model, correspond to other aspects discussed in reviews (*breakfast*, *prices*, *noise*), and as it was previously shown in Titov and McDonald (2008), aspects for global topics correspond to the types of reviewed items (*hotels in Russia*, *Paris hotels*) or background words.

Notice though, that the 3rd local topic induced for the rating *rooms* is slightly narrow. This can be explained by the fact that the aspect *rooms* is a central aspect of hotel reviews. A very significant fraction of text in every review can be thought of as a part of the aspect *rooms*. These portions of reviews discuss different coherent sub-aspects related to the aspect *rooms*, e.g., the previously discovered topic *noise*. Therefore, it is natural to associate several topics to such central aspects. To test this we varied the number of topics associated with the sentiment predictor for the aspect *rooms*. Top words for resulting topics are presented in Table 2. It can be observed that the topic model discovered appropriate topics while the number of topics was below 4. With 4 topics a semantically unrelated topic (*check-in/arrival*) is induced. Manual selection of the number of topics is undesirable, but this problem can be potentially tackled with Dirichlet Process priors or a topic split criterion based on the accuracy of the sentiment predictor in the MAS model. We found that both *service* and *location* did not benefit by the assignment of additional topics to their sentiment rating models.

The experimental results suggest that the MAS model is reliable in the discovery of topics corresponding to the rated aspects. In the next section we will show that the induced topics can be used to accurately extract fragments for each aspect.

3.2 Sentence Labeling

A primary advantage of MAS over unsupervised models, such as MG-LDA or clustering, is that topics are linked to a rated aspect, i.e., we know exactly which topics model which aspects. As a result, these topics can be directly used to extract textual mentions that are relevant for an aspect. To test this, we hand labeled 779 random sentences from the dataset considered in the previous set of experiments. The sentences were labeled with one or more aspects. Among them, 164, 176 and 263 sentences were labeled as related to aspects *service*, *location* and *rooms*, respectively. The remaining sentences were not relevant to any of the rated aspects.

We compared two models. The first model uses the first three topics of MAS to extract relevant mentions based on the probability of that topic/aspect being present in the sentence. To obtain these probabilities we used estimators based on the proportion of words in the sentence assigned to an aspects' topic and normalized within local topics. To improve the reliability of the estimator we produced 100 samples for each document while keeping assignments of the topics to all other words in the collection fixed. The probability estimates were then obtained by averaging over these samples. We did not perform any model selection on the basis of the hand-labeled data, and tested only a single model of each type.

For the second model we trained a maximum entropy classifier, one per each aspect, using 10-fold cross validation and unigram/bigram features. Note that this is a *supervised* system and as such represents an upper-bound in performance one might expect when comparing an unsupervised model such as MAS. We chose this comparison to demonstrate that our model can find relevant text mentions with high accuracy relative to a supervised model. It is difficult to compare our model to other unsupervised systems such as MG-LDA or LDA. Again, this is because those systems have no mechanism for directly correlating topics or clusters to corresponding aspects, highlighting the benefit of MAS.

The resulting precision-recall curves for the aspects *service*, *location* and *rooms* are presented in Figure 4. In Figure 4c, we varied the number of topics associated with the aspect *rooms*.[10] The average precision we obtained (the standard measure proportional to the area under the curve) is 75.8%, 85.5% for aspects *service* and *location*, respectively. For the aspect *rooms* these scores are equal to 75.0%, 74.5%, 87.6%, 79.8% with 1–4 topics per aspect, respectively. The logistic regression models achieve 80.8%, 94.0% and 88.3% for the aspects *service*, *location* and *rooms*. We can observe that the topic model, which does not use any explicitly aspect-labeled text, achieves accuracies lower than, but comparable to a supervised model.

4 Related Work

There is a growing body of work on summarizing sentiment by extracting and aggregating sentiment over ratable aspects and providing corresponding textual evidence. Text excerpts are usually extracted through string matching (Hu and Liu, 2004a; Popescu and Etzioni, 2005), sentence clustering (Gamon et al., 2005), or through topic models (Mei et al., 2007; Titov and McDonald, 2008). String extraction methods are limited to fine-grained aspects whereas clustering and topic model approaches must resort to ad-hoc means of labeling clusters or topics. However, this is the first work we are aware of that uses a pre-defined set of aspects plus an associated signal to learn a mapping from text to an aspect for the purpose of extraction.

A closely related model to ours is that of Mei et al. (2007) which performs joint topic and sentiment modeling of collections. Our model differs from theirs in many respects: Mei et al. only model sentiment predictions for the entire document and not on the aspect level; They treat sentiment predictions as unobserved variables, whereas we treat them as observed signals that help to guide the creation of topics; They model co-occurrences solely on the document level, whereas our model is based on MG-LDA and models both local and global contexts.

Recently, Blei and McAuliffe (2008) proposed an approach for joint sentiment and topic modeling that can be viewed as a supervised LDA (sLDA) model that tries to infer topics appropriate for use in a given classification or regression problem. MAS and sLDA are similar in that both use sentiment predictions as an observed signal that is predicted by the model. However, Blei et al. do not consider multi-aspect ranking or look at co-occurrences beyond the document level, both of which are central to our model. Parallel to this study Branavan et al. (2008) also showed that joint models of text and user annotations benefit extractive summarization. In particular, they used signals from pros-cons lists whereas our models use aspect rating signals.

5 Conclusions

In this paper we presented a joint model of text and aspect ratings for extracting text to be displayed in sentiment summaries. The model uses aspect ratings to discover the corresponding topics and can thus extract fragments of text discussing these aspects without the need of annotated data. We demonstrated that the model indeed discovers corresponding coherent topics and achieves accuracy in sentence labeling comparable to a standard supervised model. The primary area of future work is to incorporate the model into an end-to-end sentiment summarization system in order to evaluate it at that level.

Acknowledgments

This work benefited from discussions with Sasha Blair-Goldensohn and Fernando Pereira.

[10] To improve readability we smoothed the curve for the aspect *rooms*.

References

David M. Blei and Jon D. McAuliffe. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022.

S.R.K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

G. Carenini, R. Ng, and A. Pauls. 2006. Multi-Document Summarization of Evaluative Text. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. 2005. Pulse: Mining customer opinions from free text. In *Proc. of the 6th International Symposium on Intelligent Data Analysis*, pages 121–132.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the Natural Academy of Sciences*, 101 Suppl 1:5228–5235.

T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. 2004. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*.

A. Gruber, Y. Weiss, and M. Rosen-Zvi. 2007. Hidden Topic Markov Models. In *Proceedings of the Conference on Artificial Intelligence and Statistics*.

M. Hu and B. Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM Press New York, NY, USA.

M. Hu and B. Liu. 2004b. Mining Opinion Features in Customer Reviews. In *Proceedings of Nineteenth National Conference on Artificial Intellgience*.

C. Manning and M. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, pages 171–180.

Radford Neal. 1992. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113.

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

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

B. Snyder and R. Barzilay. 2007. Multiple Aspect Ranking using the Good Grief Algorithm. In *Proceedings of the Joint Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies*, pages 300–307.

I. Titov and R. McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17h International Conference on World Wide Web*.

P. Turney. 2002. Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Hanna M. Wallach. 2006. Topic modeling; beyond bag of words. In *International Conference on Machine Learning*.

Xuerui Wang and Andrew McCallum. 2005. A note on topical n-grams. Technical Report UM-CS-2005-071, University of Massachusetts.

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the National Conference on Artificial Intelligence*.

L. Zhuang, F. Jing, and X.Y. Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM)*, pages 43–50.