World Wide Web (2017) 20:23–37
DOI 10.1007/s11280-016-0398-9

Aspect identification and ratings inference for hotel reviews

Wei Xue[1] · **Tao Li**[1] · **Naphtali Rishe**[1]

Received: 1 March 2016 / Revised: 13 May 2016 /
Accepted: 27 May 2016/ Published online: 9 June 2016

Abstract Today, a large volume of hotel reviews is available on many websites, such as TripAdvisor and Orbitz. A typical review contains an overall rating, several aspect ratings, and review text. The rating is an abstract of review in terms of numerical points. The task of aspect-based opinion summarization is to extract aspect-specific opinions hidden in the reviews which do not have aspect ratings, so that users can quickly digest them without actually reading through them. The task consists of aspect identification and aspect rating inference. Most existing studies cannot utilize aspect ratings which become increasingly abundant on review hosts. In this paper, we propose two topic models which explicitly model aspect ratings as observed variables to improve the performance of aspect rating inference on unrated reviews. The experiment results show that our approaches outperform the existing methods on the data set crawled from TripAdvisor website.

Keywords Opinion mining · Topic models · Data mining

1 Introduction

The trend that people browse hotel reviews on websites before booking encourages researchers to analyze this valuable social media data, i.e., reviews. In a typical scenario,

http://www.tripadvisor.com

http://www.orbitz.com

✉ Wei Xue
wxue004@cs.fiu.edu

Tao Li
taoli@cs.fiu.edu

Naphtali Rishe
rishe@cs.fiu.edu

[1] School of Computing and Information Sciences, Florida International University, Miami FL, USA

users write down their own opinions and rate hotels with numerical scores. Sometimes, the scores include several aspect ratings predefined by websites such as `room`, `service`, and `location`. The overall rating score expresses a general impression of the reviewer. Although people can understand how the reviewer think about the hotel at first glance, but the overall score hides a lot of details. For example, given a review with 3 stars, it is likely that the reviewer holds different attitudes towards different aspects. Without fine-grained analysis, we cannot tell whether the user express negative or positive on what aspects, because the detailed sentiments are mixed into the general overall scores. On the other hand, users usually do not have the patience to read through the review text. To this end, the aspect-specific sentiment analysis provides a good solution. There is a lot of reviews missing aspect ratings. Identifying aspect and learning more informative aspect ratings is an attractive topic in opinion mining. It helps users gain more details of each aspect easily.

Many approaches have been proposed towards aspect-based opinion mining. A comprehensive survey [14, 15] indicates that when using opinion phases, topic model based methods outperform other bag-of-words based models. In Interdependent LDA (ILDA) [14], the vocabulary of a collection of reviews is decomposed into two sets: the head terms and the modifier terms with POS Tagging processing. Each review is assumed to be made of several pairs of heads and modifiers. For example, the phrase "nice service" is parsed into a pair of the head term "service" and the modifier term "nice". The modifier term is used to infer the sentiment polarity, while the associated head terms are the features for aspect identification. The head terms do not have sentiment polarity. Both of the head terms and the modifier terms are modeled as observed variables and conditioned on the latent variables, i.e., rating variables and topic variables. In addition, it is straightforward to consider the dependencies between the rating variables generating the modifier terms and the topic variables producing the head terms, because reviews usually have different preferences across different aspects.

However, the topic models [14, 21, 22] cannot gain any benefit from the available aspect ratings associated with reviews. Aspect ratings are now very easy to be obtained from websites like TripAdvisor and Orbitz. TripAdvisor has the largest volume of reviews, which is about 225 million. Most reviews are associated with aspect ratings. The problem of traditional topic models is that they do not explicitly model the observed aspect ratings from data. Motivated by this observation, we propose two new topic models which can simultaneously learn aspects and their ratings by utilizing the numerical aspect ratings. Our model can be applied to any review data set without aspect ratings. The aspect ratings are only needed for training. Specifically, our models are based on opinion phrases which are pairs of head and modifier terms. The dependencies between aspects and their ratings are captured by their latent variables. We use Gibbs sampling to estimate the parameters of the models on the training data set and use maximizing a posteriori (MAP) method to predict aspect ratings on unrated reviews.

A preliminary version of the work has been published in [23]. In this journal submission, in addition to revising and elaborating the original paper, we propose new topic model ARIH (Aspect and Rating Inference using Hotel specific aspect rating priors), which extends the prior models and achieves better experiment performance. The rest of paper is organized as follows. Section 3 formulates the problem and notation we use. Section 4 proposes our model and describes the inference methods. Section 5 shows the data, the experiments and discuss experiment results. Finally we draw the conclusion and provide future research tasks in Section 6.

2 Related work

The problem of review sentiment mining has been an attractive research topic in recent years. There are several lines of research. The early work focuses on the overall polarity detection, i.e., detecting whether a document is positive or negative. The author of [17] found that the standard machine learning techniques outperform human on the sentiment detection. Later, the problem of determining the reviewers sentiment with respect to a multi-point scale (ratings) is proposed in [16]. The problem was transformed into a multi-class text classification problem. Hidden Markov Model (HMM) is specially adapted to identify aspects and their polarity in Topic Sentiment Mixture model (TSM) [13]. Ranking methods are also used to produce numerical aspect scores [18].

In the literature, Latent Dirichlet Allocation (LDA) [3] based methods play a major role because the ability of topic detection of LDA is very suitable for multi-facet sentiment analysis on reviews. MG-LDA [19, 20] (Multi-Grain Latent Dirichlet Allocation) considers a review as a mixture of global topics and local topics. The global topics capture the properties of reviewed entities, while the local topics vary across documents to capture ratable aspects. Each word is generated from one of these topics. In their later work, the authors modeled the aspect rating as the outputs of linear regressions, and combine them into the model in the corresponding aspect. Joint sentiment/topic model (JST) [9, 10] focuses on aspect identification and ratings prediction without any rating information available. In JST, the words of reviews are determined by the latent variables of topic and sentiment. Aspect and Sentiment Unification model (ASUM) [6] further assumes all the words in one sentence are sampled from one topic and one sentiment. CFACTS model [7] combines HMM with LDA to capture the syntactic dependencies between opinion words on the sentence level. Given overall ratings, Latent Aspect Rating Analysis (LARA) [21, 22] uses a probabilistic latent regression approach to model the relationships between latent aspect ratings and overall ratings. On the other hand, the POS-Tagging technique is frequently used in the detection of aspect and sentiment. The authors of [11] categorized the words in reviews into head the terms and the modifier terms with simple POS-Tagging methods. They proposed a PLSI based model to discover aspects and predict their ratings. Interdependent LDA model (ILDA) [14] captures the bi-direction influence between latent aspects and ratings based on the preprocessing of head terms and modifier terms. Senti-Topic model with Decomposed Prior (STDP) [8] learns different distributions for topic words and sentiment words with the help of basic POS-Tagging. Similar ideas are applied to separate aspects, sentiments, and background words from the text [24].

Our models are based on opinion phrases [11], but overcome the drawback of previous models that cannot take advantage of the available aspect ratings. We consider the relationships between several factors, such as overall ratings, aspect ratings, head terms, and modifier terms.

3 Problem formulation

In this section, we introduce the aspect-based opinion task and list notations we use in our models. Formally, we define a data corpus of N review documents, denoted by $\mathcal{D} = \{x_1, x_2, \ldots, x_D\}$. Each review document x_d in the corpus is made of a sequence

of tokens. Each review x_d is associated with an overall rating r_d, which takes an integer value from 1 to $S(S = 5)$. An aspect is a predefined property of a hotel, such as `value`, `room`, `location`, and `service`. A text review expresses the reviewer's opinions on several aspects. For example, the occurrence of the word `price` indicates the review comments on aspect `value`. Each review is associated with several integer scores called ratings $\{l_1, l_2, \ldots, l_K\}$, where K is the number of aspects.

Phrase We assume each review is a set of opinion phrases f which are pairs of head and modifier terms, i.e., $f = \langle h,\ m \rangle$. In most cases, the head term h describes an aspect and the modifier term m expresses the polarity of the phrase. The basic NLP techniques like POS-Tagging are used to extract phrases from raw text for each review.

Aspect An aspect is a predefined attribute that the reviewers may comment on. It also corresponds a probabilistic word distribution over the vocabulary in the topic models, which can be learned from data.

Rating Each review contains an overall rating and several aspect ratings. The rating of each review is an integer from 1 to 5. We assume that the overall ratings are available for each review, but the aspect ratings are available only in the reviews used for training.

Review A review is represented as a bag of phrases, i.e., $x_d = \{f_1, f_2, \ldots, f_M\}$.

Problem Definition Given a collection of reviews, the main problem is to 1) identify aspects of reviews, and 2) infer the unknown aspect ratings on the unrated reviews.

4 Models

In this section, we propose two generative models to solve the aspect-based opinion mining task by incorporating observed aspect ratings. We list the notations of the models in Table 1. We assume reviews are already decomposed into head terms and modifier terms using NLP techniques [14].

4.1 Assumptions

We discuss some assumptions in modeling review text. First, our models presume a flow of generating ratings and text. The reviewer gives an overall rating based on his experience, then rates the hotel on some aspects and writes down review text. In the model of bag-of-phrases, the reviewer chooses a head term for an aspect on which he would like to comment, then he picks a modifier term to express his opinion. This generation process is captured by our models.

Second, the aspect ratings depend on the overall ratings. For example, when a user gives a 5-star overall rating, it is unlikely that the user gives low ratings on any of the aspects. An average overall score indicates the reviewer is disappointed on some aspects, but not all of them. It is possible that the reviewer holds positive feedbacks on other aspects. Inspired by this observation, we model the aspect ratings π with multinomial distributions $P(\pi|r)$ conditioned on the overall rating r.

Third, the aspect ratings imply another relationship with modifier terms of opinion phrases [15]. Because, for different aspects, people use different words to express different

Table 1 The table of notations

D	the number of reviews
K	the number of aspects
M	the number of opinion phrases
S	the number of distinct integers of ratings
U	the number of head terms
V	the number of modifier terms
z	the aspect / topic switcher
l	the aspect rating
h	the head term
m	the modifier term
r	the overall rating
θ	the topic distribution in a review
π	the aspect rating distribution for each topic
α	the parameter of the Dirichlet distribution for θ
β	the global aspect sentiment distribution
λ	the parameter of the Dirichlet distribution for β
δ	the parameter of the Dirichlet distribution for ϕ and ψ
ϕ	the head term distribution for each topic
ψ	the modifier term distribution for each sentiment

attitude. For example, it does not make any sense to use the word "patient" to comment on the aspect "room". We explicitly introduce random variables for modifier terms which are conditioned on aspect variables, so that meaningful aspects and sentiments can be learned from the head and the modifier terms respectively.

4.2 Motivation

Existing topic models do not require aspect ratings of reviews during model training and consider it as an advantage. It may be true in the past, since there are not many reviews containing aspect ratings. Nowadays, more review hosts, such as TripAdvisor and Orbitz, allow reviewers to rate hotels on predefined aspects. The volume of such extended reviews is growing rapidly. It is reasonable to leverage the valuable information to build more precise and accurate models. To our best of knowledge, this study is the first work to utilize the aspect ratings.

Our topic models assume aspect ratings as probabilistic variables. The aspect ratings π are scores in reviews on K aspects. They are available in the training data and hence treating them as switchers is quite straightforward. An interesting observation is the distinction between the aspect rating and the phrase sentiment. They are both sentiment switchers and are conditioned on the overall rating variable r. One is for the aspect, the other is for the phrase. If we assume that both of them are generated from the prior aspect sentiment distribution β and the overall rating r, we have ARID model (Aspect and Rating Inference with the Discrimination of aspect sentiment and phrase sentiment). The interaction between π and r is through the global β and the overall rating r. It saves the direct dependency between them. If we assume in given the aspect k, the reviewer holds the same sentiment for all the modifier terms, the discrimination between aspect sentiment and phrase sentiment becomes

redundant and can be removed. The model ARIH model (Aspect and Rating Inference using Hotel specific aspect rating priors) extends the prior model (ARIM) in our work [23]. We consider the prior probabilistic distribution β of aspect ratings for each hotel. It allows the aspect ratings of reviews to be sensitive to the characteristics of each hotel.

4.3 ARID Model

ARID model, shown in Fig. 1, captures the review generation process and the two dependencies described above. Following the conventional topic models for review analysis, we use random variables z and l to simulate the generating process of the head and the modifier terms respectively. The topic selection variable z is governed by a multinomial topic distribution θ. The sentiment variable l for each opinion phrase is determined by the aspect sentiment variables β, the overall rating r and the aspect switcher z.

Specifically, in ARID model, the variables π representing aspect ratings are shaded in the graphical representation since they are observed in the training dataset. They become latent variables for prediction on unrated reviews. The latent sentiment variable l is sampled from β_k where k is determined by the value of z. The overall rating variable r serves as a prior variable for both the aspect rating π and the phrase sentiment l.

The formal generative process of our model is as follows, where `Dir` denotes Dirichlet distribution and `Mult` denotes Multinomial distribution.

- For each aspect k and each overall rating value of r
  - Sample the aspect sentiment distribution $\beta_{r,k} \sim \texttt{Dir}(\lambda)$
- For each review x_d,
  - Sample latent topic distribution variable $\theta_d \sim \texttt{Dir}(\alpha)$
  - For each aspect k from 1 to K in the review,

    Sample aspect rating $\pi_{d,k} \sim \texttt{Mult}(\beta_{r_d,k})$
  - For each phase i from 1 to M in the review,

    Sample aspect indicator $z_i \sim \texttt{Mult}(\theta_d)$

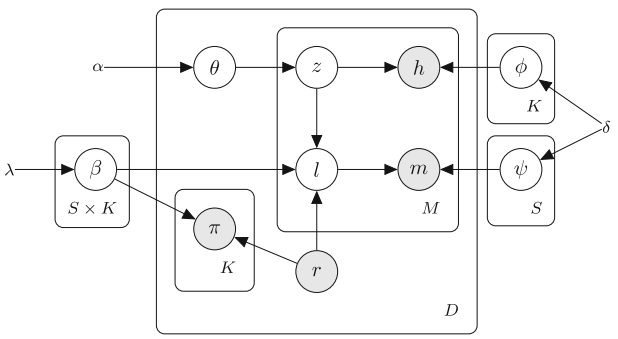

Figure 1 Graphical Representation of ARID model. The outer box represents D reviews, while the inner box contains M phrases

Sample sentiment indicator $l_i \sim \texttt{Mult}(\beta_{r_d,z_i})$
Sample head term $h_i \sim \texttt{Mult}(z_i, \phi)$
Sample modifier term $m_i \sim \texttt{Mult}(l_i, \psi)$

4.4 ARIH model

In this section, we improve a previous model. The new model ARIH (Aspect and Rating Inference using Hotel specific aspect rating priors) is shown in Fig. 2. Like the previous model ARIM (Aspect and Rating Inference Merging aspect sentiments and phrase sentiments) [23], we assume the aspect sentiment is equivalent to the phrase sentiment. In other words, the modifier terms that belong to one aspect share the same sentiment, i.e., the aspect sentiment. Therefore, we can use only one polarity indicator for both the aspect and the phrase. In particular, the aspect ratings π are modeled as in ARID, but π also indicates the phrase sentiment. Since the aspect ratings are available in the training data, the information from β to m is blocked by π according to the d-separation theory of graphical models [2]. Therefore, the modifier term is determined by the aspect ratings π instead of β, and the aspect variable z. In the generative procedure of ARIH, the modifier term m_i is sampled from $\psi_{z_i,\pi_{z_i}}$. π follows a multinomial distribution with parameter β.

In ARID and ARIM, we assume the aspect rating variables π is conditioned on the global aspect sentiment distributions β, which has the size of $K \times S$. For each aspect k and each global rating s, we have a Dirichlet distribution over ratings β. However, making the aspect rating conditioned on the global aspect sentiment distributions ignores the aspect rating biases of different hotels. Each hotel has its own pros and cons. For example, despite the hotels may have the same global rating, the one located near the airport would receive higher ratings on `location`, while those having good service may be rated higher on `service`. If both of them get 4-star ratings, the aspect ratings π have the same global prior distribution $\beta_{k,s=4}$.

To verify our assumption, we use Principle Component Analysis (PCA) to investigate the distribution of β. In particular, for each hotel, we compute the average ratings on each aspect, which give the same overall rating. It generates a matrix P, where $P_{i,j}$ is the average rating of the jth aspect of the ith hotel. We have five P matrices for each possible overall

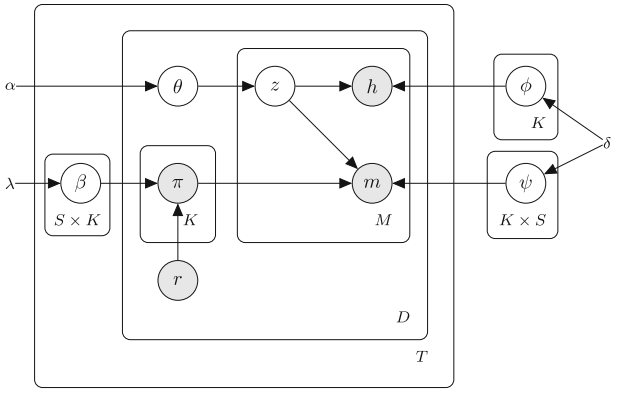

Figure 2 Graphical Representation of ARIH model

Table 2 The largest variance ratio of Principle Component Analysis

Overall rating	1	2	3	4	5
Largest variance ratio	0.704	0.753	0.831	0.871	0.965

rating ranges from 1 to 5. We use PCA to reduce the dimension of P and compute the largest variance ratio in Table 2. The variances of aspect ratings are quite large, especially for 5-star overall rating. If the variance is 0.871, for example, the ratings on some aspect can be 1-star difference for difference hotels. The analysis shows that the same overall ratings often imply different weights on aspects, which depend on the type of hotels.

ARIH associates each hotel with its own aspect rating priors $\beta_{t,r,k}$. Here, $\beta_{t,r,k}$ represents the aspect rating distribution on aspect k when the overall rating is r for hotel t. ARIH extends ARIM by using hotel-specific beta, therefore ARIM can be considered as a special case of ARIH. If we apply ARIH on the collection of reviews of one hotel. ARIH is reduced to ARIM. In Fig. 2, the graphical model ARIH has one more layer than ARIM. The variables for each review in ARIH model have one more subscript to indicate which hotel the review comes from.

4.5 Estimation

4.5.1 ARID model

There are two methods widely used for parameter estimation, i.e., Gibbs sampling [4] and variational inference [3]. Since updating equations using Gibbs sampling is relatively easy to derive and implement, we adopt collapsed Gibbs sampling (CGS) which integrates out the intermediate random variables θ, ϕ, β, and ψ. For prediction, we learn the distributions ϕ, ψ of the head and the modifier terms as well as the global aspect sentiment distribution β from z and l. The Gibbs sampling repeatedly samples latent variables $z_{a,b}$ and $l_{a,b}$ conditioned on all other latent z and l in document a for phrase b.

In ARID model, the joint probability is

$$\begin{aligned} p(z, l, h, m|\alpha, \lambda, \delta, \pi, r) = & \int p(\theta|\alpha)p(z|\theta)\times \\ & p(h|z, \phi)p(\phi|\delta)\times \\ & p(\pi|\beta, r)p(l|\beta, r, z)p(\beta|\lambda)\times \\ & p(m|l, \psi)p(\psi|\delta)\ d\theta\, d\beta\, d\phi\, d\psi\ , \end{aligned} \tag{1}$$

where we integrate out θ, ψ, β and ϕ.

We define two counters $N_{d,r,k,s,u,v}$ and $C_{d,r,k,s}$ to count the numbers of the occurrences of opinion phrases $f_{d,i} = \langle h_{d,i} = u,\ m_{d,i} = v\rangle$ and the aspect rating $\pi_{d,k}$. Specifically, $f_{d,i} = \langle h_{d,i} = u,\ m_{d,i} = v\rangle$ is the phrase i of document d which has the head term u and the modifier term v. $N_{d,r,k,s,u,v}$ is the number of times that the pair of head term u and modifier term v is assigned to aspect k and sentiment s in document d, whose overall rating is r. $C_{d,r,k,s}$ is the indicator of the document d that gives aspect rating s on aspect k when the overall rating of the document is r. Although given document d, its overall rating r_d is determined, we use the overall rating as a subscript for convenience.

$$N_{d,r,k,s,u,v} = \sum_{i=1}^{M} \mathbf{I}[r_d = r,\ z_{d,i} = k,\ l_{d,i} = s,\ h_{d,i} = u,\ m_{d,i} = v]\ , \tag{2}$$

$$C_{d,r,k,s} = \mathbf{I}[r_d = r,\ \pi_{d,k} = s]\ , \tag{3}$$

where the function **I** is the identity function. We replace the subscript N by $*$ when summing out the counter along the subscript indices. For example,

$$N_{d,r,*,s,u,v} = \sum_{k=1}^{K} N_{d,r,k,s,u,v} \ . \tag{4}$$

Gibbs sampling samples $z_{a,b}$ and $l_{a,b}$ simultaneously

$$\begin{aligned} p(z_{a,b}, l_{a,b} | z_{-(a,b)}, l_{-(a,b)}, \alpha, \delta, \lambda, h, m, r, \pi) \propto \ & (N^{-(a,b)}_{a,r_a,z_{a,b},*,*,*} + \alpha) \times \\ & \frac{N^{-(a,b)}_{*,*,z_{a,b},*,h_{a,b},*} + \delta}{N^{-(a,b)}_{*,*,z_{a,b},*,*,*} + U\delta} \times \\ & \frac{N^{-(a,b)}_{*,r_a,z_{a,b},l_{a,b},*,*} + C_{*,r_a,z_{a,b},l_{a,b}} + \lambda}{N^{-(a,b)}_{*,r_a,z_{a,b},*,*,*} + C_{*,r_a,z_{a,b},*} + S\lambda} \times \\ & \frac{N^{-(a,b)}_{*,*,*,l_{a,b},*,m_{a,b}} + \delta}{N^{-(a,b)}_{*,*,*,l_{a,b},*,*} + V\delta} \ . \end{aligned} \tag{5}$$

It turns out that the aspect ratings π can be considered as pre-observed phrase sentiment counts for the global aspect sentiment distribution β. Therefore, the prior parameter λ can be dropped. We estimate the aspect sentiment distribution β with the aspect ratings π and the overall ratings r of the training data before Gibbs sampling with (6).

$$\beta_{r,k,s} = \frac{C_{*,r,k,s}}{C_{*,r,k,*}} \ . \tag{6}$$

The third term of the right hand of (5) is replaced by

$$\frac{N^{-(a,b)}_{*,r_d,z_{a,b},l_{a,b},*,*} + \tilde{\lambda}\beta_{r_d,z_{a,b},l_{a,b}}}{N^{-(a,b)}_{*,r_d,z_{a,b},*,*,*} + \tilde{\lambda}} \ , \tag{7}$$

where $\tilde{\lambda}$ is the scaling factor for β. The parameters of ARID ψ, ϕ, θ are estimated by

$$\phi_{k,u} = \frac{N_{*,*,k,*,*,u,*} + \delta}{N_{*,*,k,*,*,*} + U\delta}, \ \psi_{s,v} = \frac{N_{*,*,*,*,s,*,v} + \delta}{N_{*,*,*,*,s,*,*} + V\delta}, \ \theta_{d,k} = \frac{N_{d,r_d,k,*,*,*,*} + \alpha}{N_{d,r_d,*,*,*,*,*} + K\alpha} \ . \tag{8}$$

4.5.2 ARIH

The iterative updating function of Gibbs sampling for ARIH has little difference from that for ARID.

$$\begin{aligned} p(z_{a,b} | z_{-(a,b)}, \alpha, \delta, \lambda, h, m, r, \pi) \propto \ & (N^{-(a,b)}_{a,r_a,z_{a,b},*,*,*} + \alpha) \times \\ & \frac{N^{-(a,b)}_{*,*,z_{a,b},*,h_{a,b},*} + \delta}{N^{-(a,b)}_{*,*,z_{a,b},*,*,*} + U\delta} \times \\ & \frac{N^{-(a,b)}_{*,*,z_{a,b},\pi_{a,z_{a,b}},*,m_{a,b}} + \delta}{N^{-(a,b)}_{*,*,z_{a,b},\pi_{a,z_{a,b}},*,*} + V\delta} \end{aligned} \ . \tag{9}$$

The parameters of ARIH model ϕ, θ is estimated by (8), but the number of ψ is $K \times S$, which is estimated by

$$\psi_{k,s,v} = \frac{N_{*,*,k,s,*,v} + \delta}{N_{*,*,k,s,*,*} + V\delta} \ . \tag{10}$$

We estimate β by

$$\beta_{t,r,k,s} = \frac{C_{t,r,k,s}}{C_{t,r,k,*}} \ , \tag{11}$$

where $C_{t,r,k,s}$ counts the number of the reviews which assign the overall rating r and the aspect rating s on the aspect k for hotel t.

4.6 Incorporating prior knowledge

We use a small set of seed words to initialize the aspect term distribution ϕ [21]. Without any prior knowledge, we have to set the number of topics and align the generated aspects with predefined aspects. It is neither necessary nor easy for analyzing hotel reviews, because we are interested in only a few widely-used aspects. We consider the seed words as the pseudo-counts, i.e., the amount of δ words are added to $\phi_{k,u}$ before Gibbs sampling.

4.7 Prediction

The goal of our models is to predict aspects and ratings on the unrated reviews. Given an opinion phrase $f_{d,i} = \langle h_{d,i}, m_{d,i} \rangle$ and the overall rating r_d in a new document d, we identify the aspect on which the phrase $\hat{z}_{d,i}$ comments and predict the aspect rating $\hat{l}_{d,i}$.

We use CGS to sample z and l together from $p(z, l|h, m, r, \alpha, \beta, \phi, \psi)$, where θ is integrated out. Here, two subscripts d and i are dropped for simplicity. After enough sampling iterations, we first estimate the predicted aspect $\hat{z}$ by the most frequent z among the pairs of $\langle z, l \rangle$. It is equivalent to use MAP (Maximum A Posterior) by integrating out l. Then given the predict $\hat{z}$, we predict $\hat{l}$ with $\mathbb{E}[p(l|\hat{z}, h, m, r, \beta, \phi, \psi, \alpha)]$. The reason why we consider the expectation of l is that the aspect ratings are numerical, rather than independent discrete category labels. The probability of each possible value l are kind of importance. The aspect mixture weight θ for a new document can be learned by Gibbs sampling as well, but we simply assume θ is a uniform distribution, because a review on hotel should comment on all the concerned aspects.

When ARIH is applied on the reviews without aspect ratings, we integrate out the latent aspect rating variable π and θ, then sample z from $p(z|h, m, r, \alpha, \beta, \phi, \psi)$ to compute MAP $\hat{z}$, like ARID model. For each opinion phrase $\langle h, m \rangle$ whose $\hat{z} = k$, we assign the most probable sentiment score $\hat{s} = \arg\max_s \psi_{\hat{z},s,m}$ to the modifier term m. Then, the estimated aspect rating $\mathbb{E}[p(\pi_k|\hat{z}, h, m, r, \beta, \phi, \psi, \alpha)]$ is computed by averaging the scores $\hat{s}$ of all the opinion phrase whose $\hat{z} = k$.

5 Experiments

In this section, we describe the review data we use and evaluate the performance of our models.

Table 3 Seed words

Aspect	Seed words
Value	value, fee, price, rating
Room	windows, room, bed, bath
Location	transportation, walk, traffic, shop
Service	waiter, breakfast, staff, reservation

Table 4 Frequentest head terms and modifier terms

Aspect	Head terms	Modifier terms
Value	deal, price, charge	good, great, reasonable
Room	house, mattress, view	comfortable, clean, nice
Location	parking, street, bus	great, good, short
Service	manager, check-in, frontdesk	friendly, good, great

5.1 Data and settings

The data set we use is crawled from TripAdvisor [21]. Each review in the data set is associated with an overall rating and 7 aspect ratings, which are within the range from 1 to 5. However some aspects such as `Cleanliness`, `Check in / front desk` are rarely rated. To better train and evaluate models, we use only four mostly commented aspects, `Value`, `Room`, `Location` and `Service`. We keep reviews with all four aspect ratings to evaluate the models. We use NLTK [1] to tokenize the review text, remove stop words, remove infrequent words, apply POS-Tagging technique [14] to extract opinion phrases, and filter out short reviews which contains less than 10 phrases. The final data set contains 1,814 hotels and 31,013 reviews. We randomly take 80 % of data as the training data set, the rest of them as the test data set. 10-fold cross validation is used to tune the hyper-parameters α and β on the training data set. The seed words used to initialize the head term distribution ϕ is in Table 3, which is a small set of words.

5.2 Aspect identification

In this section, we demonstrate that the ability of identifying meaningful aspects. Since the head terms found by the two models are not very different from each other, we present top 3 frequentest head terms for each aspect in Table 4. The listed head terms are the most frequent words, which have highest values in ϕ_k. We also list top 3 frequentest modifier terms for each aspect. The models can successfully extract ratable aspects from reviews and learn aspect-specific sentiment words as well. For example, "comfortable" is frequently used to describe aspect "Room", but not for other aspects. We also observe that people also like to use vague sentiment words for all aspects, such as "good", "great".

5.3 Metric

We use RMSE(Root-mean-square error)[1] to measure the performance of predicting aspect ratings for each hotel in the test set. Assuming the predicted aspect rating for hotel d on aspect k be $\hat{\pi}_{d,k}$ and ground-truth $\pi_{d,k}$, RMSE is represented as (12).

$$\text{RMSE}(\hat{\pi}_{d,k}, \pi_{d,k}) = \sqrt{\frac{1}{DK}\sum_{d=1}^{D}\sum_{k=1}^{K}(\hat{\pi}_{d,k} - \pi_{d,k})^2} \tag{12}$$

[1] http://en.wikipedia.org/wiki/RMSE

Table 5 Performance of Aspect Inference

Measure	Baseline	LARAM	ARID	ARIM	ARIH
RMSE	0.702	0.632	0.573	0.505	0.481
ρ_{aspect}	0.0	0.217	0.185	0.259	0.328
ρ_{hotel}	0.755	0.755	0.737	0.764	0.781

RMSE measures the accuracy of the prediction on aspect ratings. We also use Pearson correlation in (13) to describe the linear relationship between the predicted and the ground-truth aspect ratings. Here, π_d is the vector of the aspect ratings of document d.

$$\rho_{\text{aspect}} = \frac{1}{D}\sum_{d=1}^{D}\rho(\pi_d, \hat{\pi}_d) \tag{13}$$

Since the rating score is an ordinal variable, we adopt Pearson linear correlation ρ_{aspect} on the aspect ratings within each review to evaluate how a model keeps the aspect order in terms of their scores. For each aspect, we also compute the linear correlation across all hotels ρ_{hotel} as in (14). The measure is used to test whether the model can predict the order of hotels in teams of an aspect rating. π_k consists of all the aspect ratings of all the hotels on the aspect k.

$$\rho_{\text{hotel}} = \frac{1}{K}\sum_{k=1}^{K}\rho(\pi_k, \hat{\pi}_k) \tag{14}$$

5.4 Aspect rating prediction

We present the experiment results on the reviews without any aspect rating in Table 5. We compare the results between our models and one baseline. The baseline predicts all the aspect ratings of each review with the given overall ratings. The baseline predicts the aspect

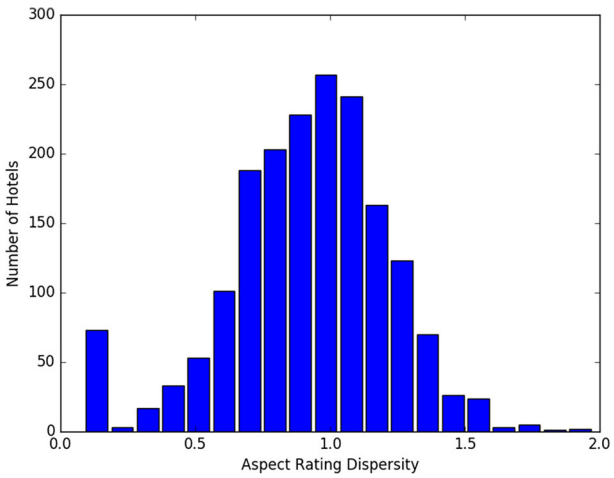

Figure 3 Aspect Rating Dispersity of Hotels

ratings of a review with a constant value, so $\rho_{\text{aspect}} = 0$. The results indicate that ARID and ARIH(ARIM) outperform the baseline and LARAM [22]. The main reason is that our models can capture the dependency between the aspects, the aspect ratings and the modifier terms, by taking into the account the aspect ratings in the training data set. In terms of ρ_{hotel}, all the approaches have similar scores. On the hotel level, the aspect ratings are averaged across all reviews, while the goals of these four methods are predicting the ratings of each individual review. The difference between each method on predicted aspect ratings for each review is small. Therefore, there is no much difference on the measure ρ_{hotel}.

Moreover, ARIH (ARIM) is better than ARID, which confirms our observation. The sentiment of aspects and modifiers is not much different from each other. Reviewers express the same polarity with different modifier terms, when commenting on one aspect. Therefore, merging aspect sentiment with modifier sentiment does not decrease the capability of the models. ARID model has K kinds of modifier term distributions ψ, while ARIH has $K \times S$, since the modifier term m in ARIH is dependent on the aspect switcher z and the aspect sentiment π. ARID estimates a global sentiment distribution across all aspects, while ARIH can learn aspect-specific sentiment distribution by modeling aspect-dependent sentiment. In the inference, the aspect on which the opinion phrases comment is determined by its head term h. ARID infers the polarity for each modifier term from a coarse sentiment distribution, while ARIH can obtain more find-grained sentiment using its $K \times S$ modifier term distributions. The parameter ψ in ARIH fine-tunes the predicting results based on β and ϕ. Therefore, in terms of Pearson correlation metric, ARIH has better performance than ARID.

ARIH model has more aspect rating distributions β than ARIM. It gives better accuracy on predicting the polarity of the modifier terms. The difference between the aspect ratings in β and those in reviews may influence the performance. Following the experiments in [12], we investigate the relationships between the dispersity and RMSE of ARIH. The dispersity is given by (15), where the we take the mean value of β and compare it with the aspect

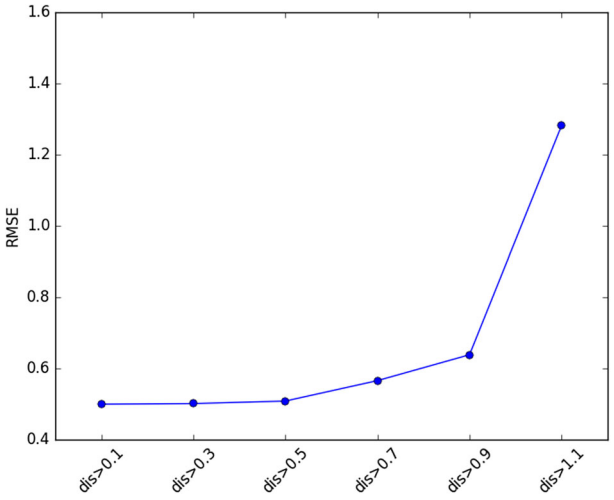

Figure 4 Aspect Rating Prediction on Reviews of Different Dispersities

ratings of reviews for each hotel. As displayed in Fig. 3, for most hotels, the aspect ratings dispersity are around 1.0 and have a clear Gaussian distribution. Moreover, there are some hotels having 0 dispersity, since the highest rating score is 5. There is very little gap between the aspect ratings and the averaged ones for highest-rated hotels.

$$\text{dis} = \sqrt{\frac{\sum_{i=1}^{K} (\mathbb{E}[\beta_{t,k}] - \pi_{t,k})^2}{K}} \tag{15}$$

As Fig. 4, we randomly prepare data set with different dispersity and report RMSE of ARIH on them. "dis >0.1" means that we use the reviews which has dispersity larger than 0.1. ARIH performs well on reviews with dispersity lower than 1.3. Due to the small training data and the high variance of aspect ratings on reviews with large dispersity, the performance of ARIH decreases.

6 Conclusion

In this paper, we propose two models for aspect identification and sentiment inference, ARID and ARIH. They utilize the overall ratings and the aspect ratings in reviews to identify the aspects and uncover the corresponding hidden aspect ratings. The models are based on topic models, but explicitly consider the dependency between the aspect ratings, the aspect terms, and sentiment terms. The opinion phrases which consist of head terms and modifier terms are extracted by simple POS-Tagging techniques. The most important contribution is that the models incorporate the aspect ratings as observed variables into the models and significantly improve the prediction performance of aspect ratings. The difference between them is that ARIH merges the sentiment variables of the modifier terms with those of the aspects. ARIH further considers the hotel-specific aspect rating priors β. Gibbs sampling and MAP is used for estimation and inference respectively. The experiments on large hotel review data set show that the models have better performance in terms of RMSE and Pearson correlation. In the future, we would investigate the methods that can automatically generate ratable aspects from the text, not from the predefined seed words. Another interesting research topic is to explore the relation between different aspects [5], because the different aspects in one review may share the similar sentiments.

Acknowledgment The work is partially supported by National Science Foundation under grants CNS-1126619, IIS-121302, and CNS-1461926 and the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

References

1. Bird, S., Klein, E., Loper, E.: Natural language processing with python. O'Reilly Media (2009)
2. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York Inc (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**(4-5), 993–1022 (2003)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. **101**(Suppl 1), 5228–5235 (2004)
5. Guo, Y., Xue, W.: Probabilistic multi-label classification with sparse feature learning. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence pp. 1373–1379

6. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the forth international conference on web search and web data mining, p. 815. ACM Press, New York, New York, USA (2011)
7. Lakkaraju, H., Bhattacharyya, C.: Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: Proceedings of the 2011 SIAM international conference on data mining, pp. 498–509 (2011)
8. Li, C., Zhang, J., Sun, J.T., Chen, Z.: Sentiment Topic Model with Decomposed Prior. In: Proceedings of the 2013 SIAM international conference on data mining. Society for industrial and applied mathematics (2013)
9. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management, p. 375. ACM Press, New York, New York, USA (2009)
10. Lin, C., He, Y., Everson, R., Ruger, S.M.: Weakly supervised joint sentiment-topic detection from text. IEEE Trans. Knowl. Data Eng. **24**(6), 1134–1145 (2012)
11. Lu, Y., Zhai, C., Sundaresan, N.: Rated aspect summarization of short comments. In: Proceedings of the 18th international conference on World wide web, p.131. ACM Press, New York, New York, USA (2009)
12. Luo, W., Zhuang, F., Cheng, X., He, Q., Shi, Z.: Ratable aspects over sentiments: Predicting ratings for unrated reviews. In: 2014 IEEE international conference on data mining, pp. 380–389 (2014)
13. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on world wide web, pp. 171–180. ACM (2007)
14. Moghaddam, S.: ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews categories and subject descriptors. In: Proceeding of the 34th international ACM SIGIR conference on research and development in information retrieval, pp. 665–674 (2011)
15. Moghaddam, S., Ester, M.: On the design of LDA models for aspect-based opinion mining. In: Proceedings of the 21st ACM international conference on information and knowledge management, pp. 803–812 (2012)
16. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the 43rd annual meeting of the association for computational linguistics, June, pp. 115–124 (2005)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02, vol. 10, pp. 79–86. Association for Computational Linguistics, Morristown, NJ, USA (2002)
18. Snyder, B., Barzilay, R.: Multiple Aspect Ranking Using the Good Grief Algorithm. In: Human language technology conference of the north american chapter of the association of computational linguistics, April, pp. 300–307 (2007)
19. Titov, I., McDonald, R.: A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th annual meeting of the association for computational linguistics, pp. 308–316. ACL (2008)
20. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th international conference on world wide web, p. 111. ACM Press, New York, New York, USA (2008)
21. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 783. ACM Press, New York, New York, USA (2010)
22. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis without aspect keyword supervision. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 618. ACM Press, New York, New York, USA (2011)
23. Xue, W., Li, T., Rishe, N.: Aspect and ratings inference with aspect ratings: Supervised generative models for mining hotel reviews. In: Web information systems engineering–WISE 2015, pp. 17–31. Springer (2015)
24. Zhao, W., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 conference on empirical methods in natural language processing, October, pp. 56–65 (2010)