Aspect-Based Opinion Polling from Customer Reviews

Jingbo Zhu, *Member*, *IEEE*, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou, *Member*, *IEEE*, and Matthew Ma, *Senior Member*, *IEEE*

Abstract—Opinion polling has been traditionally done via customer satisfaction studies in which questions are carefully designed to gather customer opinions about target products or services. This paper studies aspect-based opinion polling from unlabeled free-form textual customer reviews without requiring customers to answer any questions. First, a multi-aspect bootstrapping method is proposed to learn aspect-related terms of each aspect that are used for aspect identification. Second, an aspect-based segmentation model is proposed to segment a multi-aspect sentence into multiple single-aspect units as basic units for opinion polling. Finally, an aspect-based opinion polling algorithm is presented in detail. Experiments on real Chinese restaurant reviews demonstrated that our approach can achieve 75.5 percent accuracy in aspect-based opinion polling tasks. The proposed opinion polling method does not require labeled training data. It is thus easy to implement and can be applicable to other languages (e.g., English) or other domains such as product or movie reviews.

Index Terms—Opinion polling, sentiment analysis, opinion mining, aspect-based analysis.

1 Introduction

With the increase of opinion-rich resources such as product or movie reviews that have been publicly available, one emerging research field is opinion analysis, which relates to the study of people's assessment on social issues or products. Many recent studies on opinion analysis aimed to analyze and extract opinions or sentiment information from customer reviews and present them in the form of sentiment-based or opinion-oriented summarization [16], [17], [28], [44], [21]. In some applications, such as the film industry, people may care more about public ratings on their products, i.e., the proportion of people expressing positive opinions. This is a representative application of *opinion polling*, which gives quantitative indications of user's positive or negative opinions on products or business.

The goal of opinion polling (customer survey) is to discover customer satisfaction on a particular product, service, or business. This is traditionally done by carefully designing some questions for customers to answer. The drawbacks of such a structured survey are the expense and difficulty of question design and lack of participation because many customers do not like to participate in a question-based structured survey. To get around these difficulties, this paper focuses on opinion polling from free-form textual customer reviews, without requiring designing a set of questions in the form of a survey.

Sometimes people express their positive and negative opinions explicitly in terms of ratings, which can be easily converted into an opinion poll. However, nowadays people increasingly express their opinions in free-form textual reviews without assigning any ratings. To analyze textual reviews, some previous studies [26], [36], [32] attempted to predict the polarities of these user reviews by using supervised document classification algorithms. Some recent work [34] has expanded polarity analysis on a multipoint scale under ranking or ordinal regression frameworks in the fashion of supervised learning.

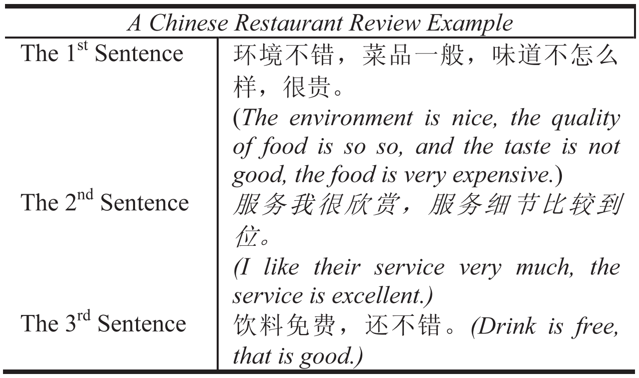
However, several challenges exist for these aforementioned approaches. First, supervised machine learning approaches are typically used and these approaches generally require labeled training data. Whereas building sufficient labeled data is often expensive and time consuming, it is desirable to develop a learning algorithm that does not require large amounts of labeled training data. Second, traditional document-level classification techniques may not always produce meaningful aspect-based opinion polling in many cases. For example, to investigate public opinions on a restaurant review *"The quality of food is so so,"* it is more important to identify some particular aspects such as *food* than overall ratings. In fact, this exemplary review expresses a negative opinion on the *food* aspect. Therefore, aspect-based opinion analysis techniques [37] are needed.

One representative work of such techniques is *feature-based sentiment summarization* [16], [17]. However, in practice, there are some limitations to using this technique for the aspect-based opinion polling addressed in this paper. First, this technique is implemented on the sentence level instead of the document level, and does not address the *polarity conflicting* issue of document-level aspect-based

- J. Zhu, H. Wang, and M. Zhu are with the Key Laboratory of Medical Image Computing (Ministry of Education), the Institute of Computer Software, College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, China. E-mail: {zhujingbo, wanghuizhen}@mail.neu.edu.cn, zhumuhua@gmail.com.
- B.K. Tsou is with the Research Centre on Linguistics and Language Information Sciences, Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong. E-mail: btsou@ied.edu.hk.
- M. Ma is with Scientific Works, 6 Tiffany Court, Princeton Junction, NJ 08550. E-mail: mattma@ieee.org.

Manuscript received 10 May 2010; revised 1 Oct. 2010; accepted 4 Jan. 2011; published online 28 Jan. 2011.
Recommended for acceptance by C.-H. Wu.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2010-05-0033.
Digital Object Identifier no. 10.1109/T-AFFC.2011.2.

1949-3045/11/$26.00 © 2011 IEEE Published by the IEEE Computer Society



A Chinese Restaurant Review Example	
The 1st Sentence	环境不错，菜品一般，味道不怎么样，很贵。 (*The environment is nice, the quality of food is so so, and the taste is not good, the food is very expensive.*)
The 2nd Sentence	服务我很欣赏，服务细节比较到位。 *(I like their service very much, the service is excellent.)*
The 3rd Sentence	饮料免费，还不错。*(Drink is free, that is good.)*

Fig. 1. A Chinese restaurant review example of three sentences.

TABLE 1
A Generated Restaurant Opinion Poll

Opinion Poll Form			
Aspect	Polarity		
	Positive (+)	Neutral (*)	Negative (-)
Environment	√		
Discount Policy	√		
Food			√
Charge			√
Service	√		

opinion analysis. Second, this technique could not satisfactorily deal with the implicit aspect expression problem that occurs frequently in restaurant reviews. Third, Hu and Liu [16], [17] used stemming, fuzzy matching, and WordNet-based synonym finding techniques to deal with the problem of word variants and misspellings for product feature generalization. In reality, various words (e.g., nouns, verbs, adjectives, and adverbs) and multiword terms can be aspect-related terms (ARTs) of the same aspect. For example, in restaurant reviews, "*noisy*," "*luxury*," and "*decoration*" are terms that all refer to the same environment aspect. Hu and Liu 's [16], [17] techniques cannot effectively tackle this problem.

Another challenge of aspect-based opinion polling lies where people often express differing opinions on multiple aspects simultaneously in the same review and even in the same sentence—this is called a *multi-aspect sentence*. For example, a real Chinese review sentence "鱼很不错，但很贵。(*the fish is great, but the food is very expensive*)" contains two different aspect mentions, including a positive food aspect mention "鱼很不错/ *The fish is great*," and a negative charge aspect mention "很贵 (*The food is very expensive*)." Therefore, treating a multi-aspect sentence as a single-aspect mention for aspect-based opinion polling would not lead to satisfactory results. In an English restaurant review set used by Snyder and Barzilay [34], more than 7 percent of sentences are multi-aspect sentences. From the Chinese restaurant reviews used in our evaluation, we find that more than 20 percent of review sentences exhibit more than one aspect. This raises an important and practical question of how to segment a multi-aspect sentence into multiple single-aspect units, referred to as *aspect-based sentence segmentation*. To our knowledge, the issue of aspect-based sentence segmentation is seldom discussed in previous studies on aspect-based opinion analysis.

To address these challenges, this paper explores the problem of aspect-based opinion polling from unlabeled free-form textual customer reviews. The key contribution of this work is threefold. First, a multi-aspect bootstrapping (MAB) method is presented to learn from unlabeled data aspect-related terms of each aspect to be used for aspect identification. Second, an aspect-based sentence segmentation model is proposed to address the challenge of segmenting a multi-aspect review sentence into multiple single-aspect units. Finally, an aspect-based opinion polling algorithm is presented in detail. All proposed methods are tested on a large real Chinese restaurant customer review data set. These methods are easy to implement, and can be easily applied to other languages such as English or other review domains such as product or movie.

2 Task Definition and Challenges

An aspect-based opinion polling system takes as input a set of textual reviews and some predefined aspects, and identifies the polarity of each aspect from each review to produce an opinion poll. Here, we first provide a real Chinese restaurant review sample, as shown in Fig. 1, to describe the aspect-based opinion polling task. An aspect-based opinion poll generated from this sample is shown in Table 1—it contains five aspects, including *environment*, *discount policy*, *food*, *charge*, and *service*, each with three polarities (i.e., *positive*, *negative*, and *neutral*).

Formally, an aspect-based opinion poll Ψ can be defined with three parameters:

- A. A finite set of aspects (e.g., *food* and *service*).
- P. A finite set of polarities, including *positive*, *negative*, and *neutral*.
- ϕ. The score function giving a voting score for each pair $(a, p) \in A \times P$.

Without loss of generality, in an aspect-based opinion polling task, we are given a set of customer reviews, $X = \{x_1, x_2, \dots, x_n\}$. The goal of aspect-based opinion polling is to calculate $\Phi(A \times P|X)$. For each pair $(a, p) \in A \times P$, we can calculate

$$\Phi(a, p|X) = \frac{\sum_{1 \le i \le n} \Phi(a, p|x_i)}{n}, \tag{1}$$

where $\Phi(a, p|x_i)$ indicates the voting score of pair (a, p) expressed by review x_i, defined as

$$\Phi(a, p|x_i) = \begin{cases} 1, & \text{if } x_i \text{ expresses polarity } p \text{ on aspect } a, \\ 0, & otherwise. \end{cases}$$

As shown in Table 1, the sample review x in Fig. 1 expresses a negative opinion on *food* aspect, namely, $\Phi\ (food, negative|x) = 1$, $\Phi(food, positive|x) = 0$, and $\Phi(food, neutral|x) = 0$. For notational convenience, we write

$\Phi(A \times P|X)$, $\Phi(a,p|X)$, and $\Phi(a,p|x_i)$ as $\Phi_\text{i}(A \times P)$, $\Phi(a,p)$, and $\Phi_\text{i}(a,p)$, respectively, for short in the following sections.

The key step in aspect-based opinion polling is to calculate $\Phi_\text{i}(a,p)$ for each review x_i. Since a real review x_i could simultaneously express multiple aspects such as *food* and *service*, a natural solution is to first analyze each review sentence in x_i and identify multiple aspects and the polarity of each aspect. Then, the overall polarity of review x_i is analyzed based on the resulting polarity of each sentence on aspect a. However, there are some crucial challenges in practice. First, casting aspect identification and polarity analysis as an n-ary classification problem is a well-studied area if sufficient labeled data are provided [26], [13], [36], [32]. Whereas it is often expensive to build labeled training data in practice, it is worthwhile studying learning algorithms for aspect identification and polarity analysis without requiring a large labeled training data set. Second, as mentioned before, considering a multi-aspect sentence as a single-aspect mention for opinion polling may not yield satisfactory results. For example, the first review sentence in Fig. 1 expresses differing opinions on three aspects, that is, positive on the *environment* aspect and negative on the *food* and *charge* aspects, respectively. A crucial issue then is how to determine if a review sentence contains multiple aspects and how to split a multi-aspect sentence into multiple single-aspect units. The third challenge is the polarity conflict problem, in which the same aspect may be associated with more than one polarity within the same review. For example, one sentence in a review represents positive on the *food* aspect and another sentence in the same review represents negative on the same aspect. Therefore, solving polarity conflict problem is a necessary step in any aspect-based opinion polling tasks.

3 Bootstrapping Aspect-Related Terms for Aspect Identification

As discussed before, the first step toward calculating $\Phi_\text{i}(a,p)$ is to identify the aspect expressed by each sentence in review x_i. Aspect identification has been well studied by using supervised learning methods for document classification or by using user-generated ontologies [13], [37]. In practice, the human labor that is associated with building a large labeled data set or domain ontologies for supervised aspect identification is often tedious and expensive. Since unlabeled data are generally publicly available and plentiful, we are led to study learning algorithms that can use unlabeled data. However, in our opinion polling task, unlabeled data alone cannot be directly used by a learning algorithm. Unlabeled data must be associated with some information or knowledge about the target function for learning tasks (i.e., aspect identification) before it can be utilized. To avoid tedious human annotation, this target information or knowledge can be given in various forms of keywords or a small number of labeled samples for training. Some researchers [33], [39], [41] have applied bootstrapping techniques to learn subjective nouns and sentiment patterns for opinionated sentence identification (i.e., subjective or objective) and polarity analysis (positive or negative). In this paper, we study the bootstrapping framework [40] for aspect identification. Seed information is used, in the form of a small number of keywords associated with aspects, to learn *aspect-related terms* (ARTs) of each aspect that are later used for aspect identification. This is further described in detail in the remaining of this section.

3.1 Aspect-Related Terms

We consider two types of ARTs in this study. First, *nouns*, *verbs*, *adjectives*, and *adverbs* are considered as the first type of candidate ARTs. These are similar to word-type features used in previous studies [14], [26], [33]. Second, some previous studies [7], [32] reported that higher order n-grams such as bigrams or trigrams in some sentiment classification settings may outperform unigrams. This motivates us to consider *multiword terms* as the second type of candidate ARTs.

To extract meaningful multiword terms from unlabeled reviews, we utilize *C-value* method [12]. Given a review set, a list of multiword terms is produced and ranked in the descending order of *C-value score*. C-value is a popular method for multiword expression extraction [12], [20], [23]. The linguistic filter used by C-value method is described as $(noun|verb|adjective|adverb)^*$. The C-value score of a multiword term t can be calculated by:

1. If t is not contained by any other terms,

$$\text{C-value}(t) = \log(|t|) \times frq(t).$$

2. Otherwise,

$$\text{C-value}(t) = \log(|t|)(frq(t) - \frac{1}{n(S)}\sum_{s\in S} frq(s)),$$

where $|t|$ denotes the number of words contained in t, $frq(t)$ indicates the frequency of occurrence of t in the corpus, S is the set of multiword terms containing t, and $n(S)$ denotes the number of terms in S.

3.2 Multi-Aspect Bootstrapping

Here, we focus on how to learn an ART set for each aspect using a *bootstrapping* algorithm that starts learning with a small number of seeds with the help of unlabeled data. Bootstrapping can be viewed as iterative clustering where, in each learning cycle, the most valuable candidate is chosen to augment the current seed set and the learning procedure continues until the predefined stopping criterion is satisfied. The key is how to measure the value score of each candidate in each learning cycle. We score each candidate ART t with the *RlogF* metric (Riloff, 1996), defined as

$$\text{RlogF}(t) = \log frq(t,T) \times R(t,T), \tag{2}$$

where T is the current seed set, $frq(t,T)$ is the frequency of co-occurrence of t and T within a given limited context (i.e., k words to left or right of t), $frq(t)$ is the frequency of co-occurrence of t in the corpus, and $R(t,T) = frq(t,T)/frq(t)$.

The proposed general bootstrapping framework for aspect-related term learning consists of the following steps, as shown in Fig. 2.

```
Input: a small number of hand-chosen seed aspect-related terms, namely $S=\{t_1, t_2, ..., t_m\}$ for aspect $a$, and unlabeled data set $U$.
● Candidate ART Extraction: Given $U$, extract nouns, verbs, adjectives, adverbs and top-$n$ multi-word terms recognized by C-value method to form a candidate ART set $\Omega$ for learning.
● Iterative Bootstrapping:
    ■ Calculate the value score RlogF for each candidate ART with respect to $S$;
    ■ Select the candidate with the highest RlogF score to augment $S$, and remove it from $\Omega$;
  Until a desirable number of ARTs have been learned.
Output: a ART set $S$ for aspect $a$.
```

Fig. 2. The general bootstrapping framework for aspect-related term learning.

Typically, we can apply this general bootstrapping algorithm to learn an ART set for each aspect independently. This technique is referred to as *single-aspect bootstrapping* (SAB). To use these learned ARTs for aspect identification, we need to assign to each learned ART an importance score that indicates the ability in reflecting the corresponding aspect. To estimate the indicative ability of each learned ART, our first intuition is to use its corresponding *RLogF* value. However, the *RlogF* value of each candidate ART, as shown in (2), is estimated using the current seed set T that is changing dynamically during the bootstrapping procedure. Therefore, the *RLogF* values of two different ARTs learned at different learning iterations are not comparable.

Let $S_i = \{t_{il}, t_{i2}, \dots, t_{ik}\}$ be an ART set for aspect a_i produced by SAB, and t_{ij} be learned at the jth learning iteration. We refer to j as the rank of term t_{ij} for aspect a_i. Bootstrapping algorithms tend to learn the most valuable ARTs for each aspect at earlier learning iterations. Therefore, a lower rank indicates that a learned ART can be more valuable for its corresponding aspect. Consequently, the indicative ability of a learned ART can be calculated in terms of its corresponding rank. In this paper, we assign a rank-based score to each ART that is learned by SAB to indicate the degree of its ability in reflecting the corresponding aspect, which is defined as

$$\eta_i(t) = 1 - \frac{r_i(t)}{|S_i|}, \tag{3}$$

where $S_i = \{t_{i1}, t_{i2}, \dots, t_{ik}\}$ is the ART set of aspect a_i learned by SAB. Notice that t_{ij} is learned in the jth iteration, $|S_i|$ indicates the number of ARTs in S_i, and $r_i(t)$ represents the rank of t in S_i, indicating in which iteration it was learned. A higher $\eta_i(t)$ value indicates that t is a more valuable ART for aspect a_i from aspect identification point of view.

For each aspect, many learned ARTs from SAB belong to more than one aspect, referred to as *multi-aspect ARTs*. For example, A, as shown in Table 2, is a multi-aspect ART.

TABLE 2
A Multi-Aspect ART A
with Respect to Three Aspects $\{a_1, a_2, a_3\}$

Rank	a_1	a_2	a_3
1			
2	$t_{1,2}$=A		
3		$t_{2,3}$=A	
4	$t_{1,4}$=B		
5			$t_{3,5}$=A
…	…	…	…

Table 2 depicts an example of multi-aspect ART A with respect to aspects a_1, a_2, and a_3. A has a lower rank than B, indicating it is more valuable for aspect a_1. However, since A is also a strong indicator for other two aspects a_2 and a_3, it actually has a weak discriminative ability for distinguishing aspect a_1 from other two aspects a_2 and a_3. We prefer single-aspect ARTs (e.g., B) to multi-aspect ARTs (e.g., A) for aspect identification because multi-aspect ARTs can be ambiguous and their importance scores should therefore be penalized. As shown in Table 2, A has a lower rank (a higher $\eta_1(A)$ value) than that of B for aspect a_1. Since A is a multi-aspect ART, its importance score should be penalized for each aspect, whereas the importance score of B for each aspect should be increased.

To address this penalty issue, we assess the ambiguity degree of each ART with respect to aspects. The ambiguity degree $\varphi(t)$ of a multi-aspect ART t can be measured by means of an entropy-like function of ranks of t with respect to aspects, defined as

$$\varphi(t) = \frac{-\sum_{i=1}^{m} \frac{r_i(t)}{\sum_{1\le j\le m} r_j(t)} \log \frac{r_i(t)}{\sum_{1\le j\le m} r_j(t)}}{\log m}, \tag{4}$$

where $S = \{S_1, S_2, \dots, S_m\}$ represents ART sets for m aspects generated by SAB. The denominator $\log(m)$ is used for normalization. A higher $\varphi(t)$ value indicates that the ART t is more ambiguous with respect to aspects.

In this paper, we present a new bootstrapping algorithm, referred to as *multi-aspect bootstrapping*, which uses an ambiguity degree factor to adjust the rank-based importance score of each multi-aspect ART. Our MAB algorithm favors the ART with a high $\eta(t)$ score and a low $\varphi(t)$ value for learning, in which the importance score assigned to an ART t for aspect a_i can be calculated by

$$score_i(t) = \eta_i(t) \times (1 - \varphi(t)). \tag{5}$$

SAB is a special case of MAB when the $\varphi(t)$ value for each ART t is set to zero. The MAB algorithm can be summarized as shown in Fig. 3.

3.3 ART-Based Aspect Identification

Without loss of generality, in the aspect identification task, we are given a sentence $C = t_1 t_2 \dots t_n$ that consists of n terms, and ART sets $S = \{S_1, S_2, \dots, S_m\}$ for m aspects. Each ART is associated with a score vector $\{score_1, score_2, \dots, score_m\}$ in which $score_i$ is its importance score for aspect a_i, defined by (5). In ART-based aspect identification, the most likely aspect a^* of sentence C is given by

```
Algorithm 2: Multi-Aspect Bootstrapping (MAB)
Input: initial aspect seed sets $S=\{S_1, S_2, ..., S_m\}$ for $m$ aspects, and unlabeled data set $U$
    ● Single-aspect bootstrapping ARTs for each aspect $a_i$, as shown in Fig. 2;
    ● Calculate the ambiguity degree of each candidate ART (Equation (4)) with respect to aspect;
    ● For each aspect, re-sort learned ARTs in the descending order of their importance scores (Equation (5)).
Output: Final generated ART sets $S^*$ for $m$ aspects.
```

Fig. 3. The multi-aspect bootstrapping algorithm for ART learning.

$$\begin{aligned} a^* &= \arg\max_{i\in[1,m]} score_i(C) \\ &= \arg\max_{i\in[1,m]} \sum_{t\in C} score_i(t). \end{aligned} \tag{6}$$

In an aspect-based opinion polling task, if the review sentence C does not contain any ARTs, we consider the most likely aspect a^* as NULL.

4 Aspect-Based Sentence Segmentation

For a review sentence that contains multiple aspects, one of the key issues for aspect-based opinion polling is to split such a multi-aspect sentence into multiple single-aspect units as the basis for aspect-based opinion polling. To tackle this problem, we propose a *multi-aspect segmentation* (MAS) model that takes a multi-aspect sentence as input and produces multiple single-aspect segments. A segment might be a subsentence,[1] or a combination of some consecutive subsentences. Let $C = c_1c_2\dots c_n$ be a sentence consisting of n subsentences, and $U = u_1u_2\dots u_k$ be its segmentation consisting of k segments. Our goal is to find the most likely segmentation U^* of the input sentence C by determining aspect changes between subsentences. Each segment expresses a particular aspect, while contiguous segments exhibit different aspects. For example, in an aspect-based sentence segmentation task, the first review sentence shown in Fig. 1 can be segmented into three single-aspect units, as shown in Fig. 4.

Sentence:
环境不错，菜品一般，味道不怎么样，很贵。(*The environment is nice, the quality of food is so so, the taste is not good, and the food is very expensive.*)
Segmentation:
环境不错 (*The environment is nice*)/ENVIRONMENT-segment || 菜品一般，味道不怎么样 (*the quality of food is so so, the taste is not good*)/FOOD-segment || 很贵 (*the food is very expensive*)/CHARGE-segment

Fig. 4. The best segmentation of the first review sentence in Fig. 1.

Our first intuition is to treat multi-aspect segmentation as a problem of traditional linear text segmentation by assuming that a sentence is a text and an aspect is a subject. While state-of-the-art linear text segmentation techniques such as *dotplotting* [5], [30], *C99* [4], and the *Fragkou method* [11] exist, practically, there are two challenges for applying these techniques to multi-aspect segmentation. First, previous studies [30], [4], [11] reported that linear text segmentation methods can work well at document level rather than sentence level because a single sentence cannot provide sufficient context to determine topic changes. Second, linear text segmentation techniques consider topic changes occurring in input text, whereas multi-aspect segmentation models deal with aspect changes occurring in the same review sentence.

It seems that an appealing solution is to incorporate aspect information of each subsentence into the design of multi-aspect segmentation models. Along this line of thinking, we formulate a multi-aspect segmentation model by introducing a criterion function $J(.)$ that aims to evaluate each candidate segmentation U of sentence C, that is,

$$U^* \overset{def}{=} \arg\max_{U} J(C,U). \tag{7}$$

The goal of this model is to find the most likely segmentation U^* with maximum $J(.)$ score. The key is how to design an appropriate criterion function $J(.)$ by incorporating aspect information of each subsentence. In the most likely segmentation U^*, two adjacent segments should express two different aspects with each segment having only one aspect. To design an appropriate criterion function $J(C,U)$ for scoring, we can utilize two types of information from segmentation U: 1) what the aspect of each candidate segment is and 2) whether the aspects of any two adjacent candidate segments are the same. It is straightforward to answer the first question, which is related to the problem of aspect identification on each segment. To determine whether two adjacent segments express the same aspect, we adopt an indicator function $\delta(u_i,u_j)$ whose value is 1 if two segments u_i and u_j are labeled as two different aspects, and 0, otherwise. Based on ART-based aspect identification techniques, the criterion function $J(C,U)$ can be formulated as

$$\begin{aligned} J(C,U) &= \sum_{1\le i\le k} [\delta(u_{i-1},u_i) \times score_{a^*}(u_i)] \\ &= \sum_{1\le i\le k} \left[\delta(u_{i-1},u_i) \times \sum_{t\in u_i} score_{a^*}(t)\right], \end{aligned} \tag{8}$$

where a^* is the most likely aspect of segment u_i and $\delta(u_0,u_1)$ is set to 1.

Let us revisit the first example sentence C in Fig. 1, which consists of four subsentences (separated by the comma),

C	a*	$score_{a*}(c_i)$
c1	Environment	0.8
c2	Food	1.5
c3	Food	0.7
c4	Charge	0.9

Candidate segmentations and *J(C,U)* scores:
$U_1=c_1c_2c_3c_4$(2.2) $U_2=c_1||c_2c_3c_4$(3.0)
$U_3=c_1c_2||c_3c_4$(2.4) $U_4=c_1c_2c_3||c_4$(3.1)
$U_5=c_1||c_2||c_3c_4$(3.2) $U_6=c_1||c_2c_3||c_4$(3.9)
$U_7=c_1c_2||c_3||c_4$(2.4) $U_8=c_1||c_2||c_3||c_4$(3.2)

Resulting segmentation $U^*=U_6=c_1||c_2c_3||c_4$(3.9)

Fig. 5. Results of multi-aspect segmentation on the first sentence in Fig. 1.

[1] The separation mark between two adjacent subsentences is defined as a comma or a semicolon.

```
Algorithm 3: Multi-Aspect Segmentation (MAS)
Input: a multi-aspect sentence $C=c_1c_2...c_n$ consisting of $n$ sub-sentences
Initialize: Set $U_0$=null, $J(C, U_0)$=0.
Loop: FOR $i$ = 1, 2, …, $n$-1
          FOR $j$ = 1, …, $n$-$i$
            1) Select a new candidate segmentation position $u$;
            2) Calculate $J(C, U_{i-1}+\{u\})$ (Equation 8)
          ENDFOR
          $u^* = \arg\max_{u} J(C,U_{i-1}+\{u\})$
          IF $J(C,U_{i-1}+\{u\}) > J(C,U_{i-1})$
          THEN $U_i \leftarrow U_{i-1}+\{u^*\}$
          OTHERWISE break;
      ENDFOR
Output: The final generated segmentation $U$*.
```

Fig. 6. Multi-aspect segmentation algorithm.

denoted by $C = c_1c_2c_3c_4$ for notational convenience. The segmentation process on C is shown in Fig. 5.

In practice, the computation cost of our multi-aspect segmentation model is related to the total number of subsentences in an input sentence, namely, $O(2^n)$, where n is the number of subsentences. Each subsentence can be a possible segment. From a computational cost stand point, it would be unrealistic to find the best segmentation by enumerating and evaluating all possible segmentations. To effectively find the most likely segmentation U^* with maximum $J(.)$ score, we can utilize the *grid search* algorithm[2] with a reasonable computational complexity $O(2^n)$. The implementation of our multi-aspect segmentation algorithm can be summarized in Fig. 6.

5 Opinion Polling Generation

An aspect-based opinion poll Ψ can be summarized as a two-dimensional form filled with $3m$ (aspect, polarity) pairs and their voting scores, involving m aspects and three polarities (i.e., positive, negative, and neutral). For each (aspect, polarity) pair, its associated voting score indicates how many customer reviews have expressed it. For example, $\Phi(food, positive) = 0.65$ indicates that 65 percent of customer reviews express positive opinions on the *food* aspect. An aspect-based opinion poll Ψ with m aspects looks as shown in Table 3. The implementation of our aspect-based opinion polling algorithm is summarized in Fig. 7.

TABLE 3
An Aspect-Based Opinion Poll with m Aspects

Aspect	Opinion Poll Ψ		
	Polarity		
	Positive (+)	Neutral (*)	Negative (-)
Aspect_1	Φ(1,+)	Φ(1,*)	Φ(1,-)
Aspect_2	Φ(2,+)	Φ(2,*)	Φ(2,-)
Aspect_3	Φ(3,+)	Φ(3,*)	Φ(3,-)
…	…	…	…
Aspect_m	Φ(m,+)	Φ(m,*)	Φ(m,-)

```
Algorithm 4: Aspect-based Opinion Polling
Input: a review set $X=\{x_1, x_2, ..., x_n\}$, $m$ predefined aspects $A=\{a_1, a_2, ..., a_m\}$ and ART sets $S=\{S_1, S_2, ..., S_m\}$ for $m$ aspects.
FOR each review $x_i$
  1. Segment each multi-aspect sentence into multiple single-aspect units. Notice that a single-aspect sentence can be viewed as a single-aspect unit;
  2. Identify the aspect and the polarity of each single-aspect unit;
  3. FOR each aspect $a_j$,
       ● Calculate $\Phi_i(a_j,+)$, $\Phi_i(a_j,-)$ and $\Phi_i(a_j,*)$ with respect to review $x_i$;
     ENDFOR
ENDFOR
FOR each aspect $a_j$,
  ● Calculate $\Phi(a_j,+)$, $\Phi(a_j,-)$ and $\Phi(a_j,*)$ with respect to $X$;
ENDFOR
Output: a final generated aspect-based opinion poll $\Psi$
```

Fig. 7. Aspect-based opinion polling algorithm. "+", "-", and "*" denote the polarities of positive, negative, and neutral, respectively.

In this paper, for polarity analysis on each single-aspect textual unit, we use a readily available Chinese sentiment lexicon[3] released by *Hownet* [9] that has been widely used for polarity analysis [38]. This sentiment lexicon contains 3,730 positive words and 3,116 negative words. Thirteen negation words such as "不/not" are used to handle the negation issue. A semantic orientation value of a textual unit is computed by summing the polarity values of all sentiment words occurring in the unit. If a sentiment word w is negated, it is converted into a new token "*NOT-w*" associated with an opposite polarity. The polarity value is empirically set to $+1$ for a positive word and -1 for a negative word. The resulting semantic orientation value of a textual unit indicates its corresponding polarity, that is, > 0 for positive, < 0 for negative, and equal to 0 for neutral. In an aspect-based opinion polling task, if one aspect is not exactly expressed in the review, its corresponding polarity is considered to be neutral.

As mentioned earlier in this paper, two conflicting polarities may exist for the same aspect in the same review. To solve the polarity conflict problem, we adopt a method that determines the predominant polarity based on the majority (e.g., > 50 percent) of polarities among competing pairs expressed in the same review, as in [41]. For example, if the number of segments expressing the pair (service, positive) is greater than that of the pair (service, negative) in the same review, the predominant polarity of service aspect is considered to be positive.

The goal of opinion polling is to give quantitative indications of user's positive or negative opinions about products or business. To achieve this goal, the crucial step is to identify the polarity of each aspect expressed in each review. Our technique used to address the polarity conflict problem seems to be a little problematic in some real-world opinion polling applications because multiple possible

[2] Our grid search algorithm treats multi-aspect sentence segmentation as a multidimensional optimization problem, in which one dimension is optimized at each round. In practice, other dynamic programming techniques such as Viterbi algorithm can also be utilized.

[3] http://www.keenage.com/html/c_index.html.

TABLE 4
Distribution Analysis of Three Types of Review Sentences

Types	Number of Sentences (%)
Multi-aspect sentence	235 (23.5%)
Single-aspect sentence	509 (50.9%)
NULL-aspect sentence	256 (25.6%)
Total	1000 (100%)

polarities can be expressed by a real review. To further address this polarity conflict issue, an alternative "*soft*" approach is to study a probabilistic framework under which a nonzero probability can be assigned to each possible polarity of an aspect. In this case, $\Phi(a, p|x_i)$ in (1) should be redefined as a real-value function that indicates a real-value voting score of pair (*a*,*p*) expressed by review x_i. In this work, we consider a simple case of using a binary function $\Phi(a, p|x_i)$ defined in (1). It is noteworthy that our aspect-based opinion polling algorithm can be implemented based on a real-value function $\Phi(a, p|x_i)$. In our future work, we will further study such soft solution to address the polarity conflict issue.

6 Experimental Evaluation

6.1 Data Set Setup

We evaluate various aspect-based opinion polling methods on a corpus of real Chinese restaurant reviews collected from the *DianPing.com*. The data set contains 13,358 reviews (54,747 sentences in total) for 100 restaurants. This data set only contains textual reviews without explicit user-provided aspect ratings. In the preprocessing step, we utilized the NEUCSP[4] tool to implement Chinese word segmentation and POS tagging. We randomly selected 1,000 sentences to investigate the distribution of different types of sentences, including multi-aspect, single-aspect, and NULL-aspect sentences. Table 4 depicts that 23.5 percent of sentences express more than one aspect, namely, multi-aspect sentences. Therefore, multi-aspect segmentation is a very important practical issue for aspect-based opinion polling.

We designed some experiments to evaluate the effectiveness of various aspect-based opinion polling methods. A subset containing 3,325 randomly chosen reviews is used for this evaluation.[5] To form the gold standard, two human judges were asked to build an opinion poll table (the same as shown in Table 1) for each review which contains a set of five aspects covering *environment*, *discount policy*, *food*, *charge*, and *service*, and was manually labeled with respect to polarities (i.e., *positive*, *negative*, or *neutral*). The inter-annotator agreement between the two human annotators is 96 percent. In case of disagreements, a third human judge is asked to act as an adjudicator. The rest of the reviews in the data set (41,641 sentences in total) were used as unlabeled data for bootstrapping-based ART learning.

TABLE 5
Initial Aspect Seed Sets for Bootstrapping

Aspect	Seeds
Environment (Env)	环境/*environment*, 豪华/*luxury*, 装修/*decoration*,嘈杂/*noisy*, 吵闹/*noisy*
Discount Policy (Dis)	打折/*discount*, 免费/ *free*, 赠券/*coupon* 优惠/*on sale*, 赠送/ *gift and rebate*
Food (Foo)	食物/*food*, 口味/*taste*, 油腻/*oily*, 好吃/*delicious*, 正宗/*authentic*
Charge (Cha)	价格/*price*, 贵/*expensive*, 便宜/*cheap*, 买单/*pay the bill*, 性价比/*cost performance*
Service (Ser)	服务员/waiter, 体贴/*considerate*, 服务/service,周到/*good service* 热情/*friendly*

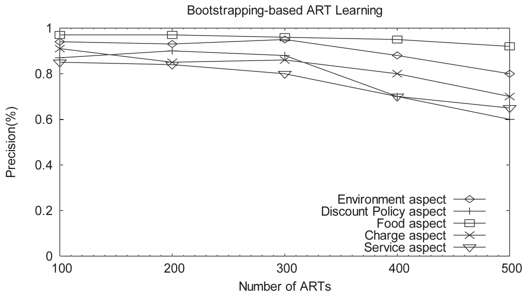


Fig. 8. Precision performance of the top-500 ARTs for each aspect learned by MAB.

6.2 Bootstrapping Aspect-Related Terms for Aspect Identification

In ART learning experiments, we considered *bigrams*, *trigrams*, and *4-grams* to be candidate multiword terms. We first extracted nouns, verbs, adjectives, and adverbs from unlabeled data, and used the C-value method to select top-40,000 multiword terms (ranked in the descending order of C-value score) to form the candidate ART set. The k value of the limited context (used in (2)) was empirically set to 5.

Selecting an appropriate number of good seeds for bootstrapping is an important step, and how to select the best seeds for bootstrapping in real-world applications is still an open question. In our solution, 500 frequently used nouns, verbs, and adjectives were first automatically extracted for manual selection of seeds. We have experimented with different numbers of seeds (i.e., 5, 10, 15, and 20) for bootstrapping, and found that the learned results are very similar after 100 iterations. Therefore, to minimize human efforts, in each bootstrapping-based ART learning algorithm, five seeds were manually chosen for each aspect, as shown in Table 5, for they appear frequently in the unlabeled corpus. In each bootstrapping algorithm, the stopping criterion is defined as when 2,000 ARTs have been learned.

To evaluate the bootstrapping results, all learned ARTs for each aspect are ranked in descending order of their importance scores. We manually checked up[6] to top-500 ARTs for each aspect produced by MAB, and reported the precision performance for each aspect as shown in Fig. 8. As

[4] NEUCSP is a Chinese word segmentation and POS tagging tool at (http://www.nlplab.com/chinese/source.htm).
[5] http://www.nlplab.com/aspect-based-opinion-poll.rar.

[6] Three annotators were asked to manually check these learned ARTs for each aspect. The same judgment decision made by at least two annotators is accepted.

TABLE 6
Some ART Examples for Each Aspect Learned by MAB

ARTs	Env	Dis	Foo	Cha	Ser
别致(unique)	0.95	0.00	0.00	0.00	0.00
包房 (KTV)	0.82	0.00	0.00	0.00	0.00
背景音乐 (background music)	0.56	0.00	0.00	0.00	0.00
安静*(peaceful)	0.51	0.00	0.00	0.11	0.15
竹笋(bamboo shoot)	0.00	0.00	0.47	0.00	0.00
中档(middle-grade)	0.00	0.00	0.00	0.69	0.00
不搭理 (ignore)	0.00	0.00	0.00	0.00	0.69
傲慢(arrogance)	0.00	0.00	0.00	0.00	0.61
菜名(menu item)	0.00	0.00	0.00	0.00	0.51
总体感觉* (overall impression)	0.09	0.00	0.00	0.07	0.08

Terms with "" indicate multi-aspect ARTs.*

can be shown, the precisions of the top-300 learned ARTs of all aspects are above 80 percent. Also, the bootstrapping algorithm works best for the *food* aspect and worst for the *discount policy* aspect. This is because, generally, more food-related terms (e.g., menu items) are mentioned in restaurant reviews, as compared to the *discount policy* aspect. From further examination of our experimental results, we found that the most valuable ARTs for the *discount policy* aspect have been learned before the 300th learning iteration, and many of those terms after the 300th iteration are noisy terms. Table 6 depicts some learned ARTs and their importance scores produced by MAB.

We further designed some experiments to investigate the effectiveness of applying these learned ARTs for aspect identification, as compared to supervised learning methods. For experimental comparison, the supervised learning method was designed based on Maximum Entropy classifier (MaxEnt) [2], a state-of-the-art classification model. The Information Gain (IG) method [19] was used for feature selection. In these experiments, 1,000 single-aspect review sentences were randomly chosen for human annotation on five aspects. We tested our ART-based aspect identification techniques and MaxEnt-based method on these sentences in terms of *accuracy*. A fivefold cross validation was performed, and the average accuracy of five trials for each method is shown in Table 7. Note that for MaxEnt-based classifier, the best accuracy was achieved by using 3,000 features.

Supervised learning methods such as MaxEnt model can achieve good performance (86 percent accuracy in our experiment) if the labeled training data are available. In general, the performance of a supervised learning method depends on the size of labeled training corpus. As illustrated in Table 7, the method based on ARTs as proposed in this paper does not use explicitly labeled aspect data, yet its performance is only 5 percent lower than that of the supervised learning method. Further, the MAB-based method outperforms the SAB-based method. In MAB, multi-aspect ARTs are penalized for their importance scores. This yields a positive effect on the performance of aspect identification.

TABLE 7
Results of Different Methods for Aspect Identification

Method	Accuracy
MaxEnt-based Classifier	0.86
Based on ARTs learned by SAB	0.77
Based on ARTs learned by MAB	0.81

6.3 Aspect-Based Sentence Segmentation

We randomly select 1,000 review sentences for evaluating the effectiveness of various aspect-based sentence segmentation algorithms. To form the gold standard, two human judges were asked to segment each review sentence in terms of aspect change (i.e., *environment, discount policy, food, charge*, and *service* aspects) and polarity change (i.e., positive and negative). That is, each multi-aspect or single-aspect and multipolarity sentence was manually segmented into multiple single-aspect and single-polarity segments, and a single-aspect and single-polarity sentence is considered as a single segment. The test set was annotated separately to verify the interannotator agreement and to verify whether the task is well defined. For disagreements between judges, a third human judge acted as an adjudicator.

The classical *precision*, *recall*, and *F1* metrics are used to evaluate the effectiveness of each automatic segmentation methods. However, the shortcoming of precision and recall for segmentation is that every inaccurately estimated segment boundary is penalized equally whether it is near or far from a true segment boundary. In the following experiments, we also adopt the *WindowDiff* metric [27] which has been widely used in text segmentation research. The WindowDiff metric is defined by Pevzner and Hearst [27]

$$\begin{aligned} &WindowDiff(ref, hyp)\\ &= \frac{1}{N-k}\sum_{i=1}^{N-k}\left(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0\right), \end{aligned}$$

where ref and hyp represent the reference segmentation and a hypothesized segment. $b(i,j)$ denotes the number of segmented boundaries between positions i and j in the sentence, and N denotes the number of subsentences in that sentence. The k value used in the WindowDiff metric is set to 3, which is the average number of subsentences per segment in the gold standard.

In our experiment, we constructed two simple baseline methods. The first baseline method is to segment a sentence by comma, named the *comma-based method*. That is, each subsentence is viewed as a single segment. The second baseline method is to simply consider the whole sentence as a single single-aspect segment named the *full-stop-based method*. In addition, we also evaluate the effectiveness of a state-of-the-art linear text segmentation method *Dotplotting* [30] and our proposed segmentation model (MAS). We used the package developed by Choi to implement the Dotplotting algorithm (www.lingware.co.uk/homepage/freddy.choi/index.htm).

The smaller the WindowDiff value, the better the segmentation performance. Among these four segmentation methods, the comma-based method is the worst. As mentioned in Table 4, only 23.5 percent of review sentences

TABLE 8
Precision (P), Recall (R), and F1 Performances of Each Method for Aspect-Based Sentence Segmentation

Methods	P	R	F1
Full-stop-based method	0.68	0.44	0.54
Comma-based method	0.17	0.38	0.24
Dotplotting	0.19	0.39	0.25
Our proposed model (MAS)	**0.69**	**0.56**	**0.62**

The bold number denotes the best performance.

TABLE 9
WindowDiff Values of Different Methods for Aspect-Based Sentence Segmentation

Methods	WindowDiff
Full-stop-based method	0.21
Comma-based method	0.72
Dotplotting	0.69
Our proposed model (MAS)	**0.17**

The bold number denotes the best performance.

in the restaurant reviews are multi-aspect cases. The comma-based method often makes wrong segmentation decisions on sentences containing one or more commas.

The dotplotting method obtains unsatisfactory performances, and its performance is very close to that of the comma-based method. The dotplotting method adopts local (i.e., between adjacent subsentences) similarity measures based on word repetitions [30]. From segmentation results, we find that many single-aspect and single-polarity sentences have been mistakenly segmented by dotplotting. In most review sentences, there are a few common words occurring in two adjacent subsentences. In such a case, the dotplotting tends to split individual sentences at commas. Experimental results show that the dotplotting method fails to identify most true segments consisting of multiple consecutive subsentences.

As shown in Tables 8 and 9, our segmentation model achieves the best performance in terms of all evaluation metrics. There are two possible reasons why our segmentation method outperforms the state-of-the-art linear text segmentation method. First, as mentioned above, for aspect-based sentence segmentation, one sentence cannot provide sufficient context for traditional linear text segmentation techniques to determine topic changes occurring in the input sentence. Second, our model explicitly utilizes aspect knowledge such as ARTs in sentence segmentation whereas linear text segmentation methods do not consider any aspect and polarity information expressed in a sentence.

6.4 Aspect-Based Opinion Polling

In the following experiments, a fivefold cross validation was performed, and the average accuracy of five trials for each process is shown in Table 10. Since the linear text segmentation method obtains unsatisfactory performance on aspect-based sentence segmentation, it was not included in this experiment for aspect-based opinion polling. We tested five different methods for aspect-based opinion polling as follows:

TABLE 10
Results of Various Automatic Aspect-Based Opinion Polling Methods

	SAB+ Full-Stop	SAB+ MAS	MAB+ Full-Stop	MAB+ MAS
Environment	0.582	0.644	0.586	0.692*
Discount Policy	0.855	0.850	0.881*	0.874
Food	0.435	0.547	0.589	0.644*
Charge	0.666	0.683	0.684	0.721*
Service	0.782	0.835	0.774	0.843*
Average	0.664	0.712	0.703	0.755*
	PRanking (Snyder and Barzilay 2007)			
Environment	0.252			
Discount Policy	0.863			
Food	0.307			
Charge	0.508			
Service	0.612			
Average	0.508			

The number with a symbol (*) indicates the best accuracy performance.

1. $SAB + MAS$ method adopts SAB for ART learning, and MAS for aspect-based sentence segmentation.
2. $SAB + \textit{Full-Stop}$ method adopts SAB for ART learning, and does not implement aspect-based sentence segmentation.
3. $MAB + MAS$ method adopts MAB for ART learning, and MAS model for aspect-based sentence segmentation.
4. $MAB + \textit{Full-Stop}$ method adopts MAB for ART learning and does not implement aspect-based sentence segmentation.
5. *PRanking* [6] is an ordinal regression model and was used for content-based rating inference in the restaurant review domain [34]. In this baseline method, we utilize the PRanking algorithm to implement aspect-based opinion polling by considering it as a rating inference problem. Along this line, to train the PRanking model, for each aspect we consider negative, neutral, and positive polarities as rating 1, 2, and 3, respectively. After the removal of stop words, each review in the data set is represented as a vector of unigram lexical features and is accompanied by a set of five aspects. For each aspect, each review is rated using a three-point scale. Following the previous work [34], the number of training iterations for PRanking algorithm is set to 4.

The effectiveness of each polling method is measured by the *accuracy*, i.e., the fraction of automatically identified (aspect, polarity) pairs that are correct.

Tables 10 and 11 shows the effectiveness of various automatic methods for aspect-based opinion poll generation. All methods except the PRanking algorithm work in an unsupervised learning fashion without requiring labeled training data. In the PRanking algorithm, aspect-based opinion polling is treated as rating inference using a three-point scale instead of polarity. The PRanking algorithm performs the worst on four out of five aspects except for the discount policy aspect, and is the worst among all methods in terms of average accuracy. As mentioned before, in

TABLE 11
Paired t-Tests between the Various Methods in Terms of *Average Accuracy*, Involving $\mathrm{SAB} + \text{Full-Stop}\ (\mathrm{S} + \mathrm{F})$, $\mathrm{SAB} + \mathrm{MAS}\ (\mathrm{S} + \mathrm{M})$, $\mathrm{MAB} + \text{Full-Stop}\ (\mathrm{M} + \mathrm{F})$, $\mathrm{MAB} + \mathrm{MAS}\ (\mathrm{M} + \mathrm{M})$, and the Method of Snyder and Barzilay (2007) (PRanking)

	S+F	S+M	M+F	M+M	PRanking
S+F	N/A	<<	<<	<<	>>
S+M	>>	N/A	~	<<	>>
M+F	>>	~	N/A	<<	>>
M+M	>>	>>	>>	N/A	>>
PRanking	<<	<<	<<	<<	N/A

Given p-value > 0.05*, the notions of A (row) "*$\gg$*" B (column), "*$\ll$*" and "*$\sim$*" indicate A is better, worse, and not significantly different than/from B on performance comparison, respectively.*

aspect-based opinion polling, the polarity neutral indicates no comment on some aspect. In such a case, the PRanking algorithm obtains unsatisfactory performance on predicting correct neutral on some aspect of no comment. For example, a test review is associated with true neutral for *food* aspect and positive for the rest. In this case, the Pranking algorithm tends to predict incorrect positive on food aspect because of positive for other four aspects.

Table 10 further depicts that $\mathrm{MAB} + \mathrm{MAS}$ achieves the highest accuracy of 75.5 percent by averaging all five aspects, followed by $\mathrm{SAB} + \mathrm{MAS}$. Therefore, methods that involve MAS achieve high accuracies than those without MAS. This is anticipated, as more than 23 percent of review sentences express multiple aspects, as shown in Table 4. The $\mathrm{MAB} + \mathrm{MAS}$ method also works the best among other methods on four out of five aspects, with the exception of the *discount policy* aspect. This is because most of the sentences involving the *discount policy* aspect are single-aspect sentences. It is further observed from real restaurant reviews that most of multi-aspect review sentences generally involve *environment*, *food*, and *service* aspects. Consequently, as shown in Table 10, our segmentation techniques (MAS) contribute to better performance on *environment*, *food*, and *service* aspects than their corresponding counterparts *Full-Stop* techniques.

When using the same segmentation method, the MAB algorithm outperforms the SAB algorithm. This is because the MAB algorithm prefers single-aspect ARTs to multi-aspect ARTs, and penalizes the importance scores of multi-aspect ARTs. Single-aspect ARTs have more discriminating power than multi-aspect ARTs, which result in a better performance in aspect identification as shown in Table 7. MAB can provide more help to aspect-based opinion polling than SAB.

Table 10 also depicts that the $\mathrm{SAB} + \mathrm{MAS}$ method achieves better average accuracy performance than the $\mathrm{MAB} + \text{full-stop}$ method. As mentioned before, the MAB-based aspect identification method outperforms the SAB-based method. However, in comparison to MAS, the full-stop-based method can result in negative effect on aspect-based opinion polling performance due to multi-aspect sentences frequently occurring in real customer reviews.

7 Related Work

Recent work of aspect-based opinion analysis has generally focused on product reviews [28], [18], [13], [22], [29], [16], [17], [8]. The most related work is *feature-based sentiment summarization* [16], [17], which aimed at producing a summary expressing the aggregated sentiment for each feature of a product and supporting textual evidence. Their system FBS performs in three main steps, including mining product features, identifying opinion sentences and deciding the semantic orientation of each opinion sentence, and summarizing the results. Their experiments demonstrated that this simple method performed well on customer reviews of five types of products such as digital cameras, players, cellular phones, routers, and software. Ding et al. [8] further proposed a holistic lexicon-based approach to improve the method of Hu and Liu [17] by addressing two issues: 1) opinion words that are content dependent,and 2) aggregating multiple opinion words in the same review sentence. However, their techniques are not adequate in dealing with aspect-based opinion polling focused in this paper.

First, implicit aspects frequently occur in restaurant reviews due to the informal writing style for online comments. For example, two frequent Chinese reviewing sentences "很贵/*too expensive*" and "太吵/*too noisy*" are implicit negative aspects for *charge* and *environment*, respectively. Hu and Liu [16], [17] only focused on mining features that explicitly appear as nouns or noun phrases in the product reviews and cannot effectively deal with the implicit feature expression problem. Second, Hu and Liu [16], [17] assumed that product features and opinion words explicitly appear as noun phrases and adjectives, respectively. Their technique cannot handle the problem that an opinion word (e.g., adjective *expensive*) can not only explicitly indicate a sentiment polarity, but also implicitly express an aspect (feature) simultaneously. For example, the word "太吵/too noisy" explicitly indicates negative and implicitly indicates environment aspect. Third, Hu and Liu [16], [17] used stemming and fuzzy matching techniques to deal with the problem of word variants and misspellings for feature generalization. For example, "*picture*," "*pictures*," and "*picture*"[7] all indicate the same digital camera feature. However, their techniques cannot effectively tackle the problem that various words (nouns, verbs, adjectives, and adverbs) and multiword terms can be aspect-related terms of the same aspect in restaurant reviews, e.g., three environment aspect terms "*noisy*," "*luxury*," and "*decoration*." Fourth, their FBS system performs on the sentence level, while aspect-based opinion polling is implemented on the document level. In such a case, their technique could not deal with the document-level polarity conflict problem that is one of the crucial issues of aspect-based opinion polling. Besides, it is common in real restaurant reviews that many review sentences express more than one aspect and are difficult to express in an aggregated sentiment. Aspect-based sentence segmentation seems to be a feasible solution to address this challenge, which is seldom addressed by Hu and Liu, or in other previous studies [13], [44]. Here, we propose an MAS

[7] It is a misspelling of "picture."

model to segment a multi-aspect sentence into multiple single-aspect units for aspect-based opinion polling. Unlike our aspect-based sentence segmentation, the method of Hu and Liu [16], [17] identified one or more features expressed by a reviewing sentence, but still used the whole sentence (instead of segments) for feature-based summarization generation.

Titov and McDonald [37] used the aspect ratings manually provided by users for the purpose of sentiment analysis, whereas in our work, the aspect-based opinion poll is generated automatically from unlabeled reviews without the user deliberately entering aspect ratings. In related areas of opinion extraction from user reviews, some previous efforts have focused on the extraction of opinion topics [16], [28] that is limited to extracting the mentions of product names and their features. Kim and Hovy [21] presented a technique for extracting opinion topics based on semantic frames, and provided a limited evaluation. However, these previous efforts did not address the issue of segmentation of multi-aspect sentences for aspect mention extraction or opinion topic extraction.

Some researchers [33], [39], [41] have applied bootstrapping frameworks for learning subjective words or sentiment patterns, in the same fashion as our single-aspect bootstrapping method. We propose a new multi-aspect bootstrapping framework for learning aspect-related terms for each aspect from unlabeled customer reviews.

Another related work was done by Thomas et al. [36]. In their work, the support/oppose sentiment classification problem was solved using a supervised SVM-based classifier. In more recent work [24], [34], polarity analysis has been expanded to a multipoint scale under a ranking or ordinal regression framework. Although these may be the most similar work to ours, they focused on supervised or semi-supervised learning techniques for sentiment analysis which need labeled data for training. However, in our work, no explicitly labeled data are needed for training.

Text segmentation is an important problem in information retrieval. Previous studies [30], [4], [11], [15] focused on text segmentation at the document level instead of the sentence level. In our comparison experiments, we evaluated a state-of-the-art linear text segmentation technique for aspect-based sentence segmentation. Results showed that the linear text segmentation technique yielded unsatisfactory performance in the aspect-based sentence segmentation task because a single sentence cannot provide sufficient context to determine topic changes for segmentation.

8 Discussion and Future Work

There are two key modules used in our aspect-based opinion polling algorithms: bootstrapping-based ART learning and multi-aspect sentence segmentation. A practical issue of bootstrapping learning is how to predefine some appropriate ART seeds manually in advance. This is a hard problem in practice. In this study, we manually selected five seeds per aspect from 500 frequent used nouns, verbs, and adjectives. It is worthy studying an effective technique to automatically or semi-automatically acquire some seeds for bootstrapping learning.

In some cases where a sentence contains multi-aspect subsentences, our aspect-based sentence segmentation model would fail. This is because our model works on the subsentence level, in which a subsentence is viewed as the basic unit (i.e., single-aspect unit) for aspect-based sentence segmentation. To solve this problem, it is worthwhile to further study the aspect-based segmentation model on the word level instead of the subsentence level.

From the experimental results, we found a *mixed opinion problem* that the number of segments expressing positive sentiments toward an aspect is approximately equal to the number of segments expressing negative sentiments in the same review. In such a case, the predominant polarity of this aspect cannot be simply considered as positive or negative. One of the alternative ways is to assign a degree score to each possible polarity of the same aspect expressed in such a review.

Domain adaptation is also important for sentiment analysis and opinion mining [1], [7]. Since most readily available sentiment lexicons are general-purpose knowledge bases, it is worth studying how to automatically adapt a general-purpose sentiment lexicon to a domain sentiment lexicon to achieve better performance. In approaching this, there are at least three crucial issues to be considered. First, the same sentiment term might indicate different polarities in different domains. Second, the difference in sentiment vocabularies across different domains should be considered. The third issue is how to assign a strength marker to each sentiment word.

9 Conclusion

This paper proposes an automatic method of aspect-based opinion polling from unlabeled textual customer reviews. A multi-aspect bootstrapping method is proposed to learn aspect-related terms of each aspect to be used for aspect identification. A multi-aspect segmentation model is proposed to handle multi-aspect sentences. Finally, an aspect-based opinion polling algorithm is presented in detail. Some evaluation experiments are designed to run on real Chinese restaurant reviews. Our approaches can be easily applied to aspect-based English language opinion analysis or in other review domains if one of the English sentiment lexicons such as *SentiWordNet*[8] is readily available for sentiment classification (polarity analysis). In our future work, we will focus on sentiment lexicon domain adaptation for sentiment analysis, and investigate effective content-based rating inference techniques for aspect-based opinion polling.

Acknowledgments

This work was supported in part by the National Science Foundation of China (60873091, 61073140), the Specialized Research Fund for the Doctoral Program of Higher Education (20100042110031), and the Fundamental Research Funds for the Central Universities.

[8] http://sentiwordnet.isti.cnr.it/.

REFERENCES

[1] A. Aue and M. Gamon, "Customizing Sentiment Classifiers to New Domains: A Case Study," *Proc. Recent Advances in Natural Language Processing,* 2005.

[2] L. Berger Adam, V.J. Della Pietra, and S.A. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics,* vol. 22, no. 1, pp. 39-71, 1996.

[3] G. Carenini, R. Ng, and A. Pauls, "Multi-Document Summarization of Evaluative Text," *Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics,* pp. 305-312, 2006.

[4] F.Y.Y. Choi, "Advances in Domain Independent Linear Text Segmentation," *Proc. First Meeting North Am. Chapter Assoc. for Computational Linguistics,* pp. 26-33, 2000.

[5] K.W. Church, "Char Align: A Program for Aligning Parallel Texts at the Character Level," *Proc. 31st Ann. Meeting Assoc. for Computational Linguistics,* pp. 1-8, 1993.

[6] K. Crammer and Y. Singer, "Pranking with Ranking," *Proc. Neural Information Processing Systems,* pp. 641-647, 2001.

[7] K. Dave, S. Lawrence, and D.M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. Int'l Conf. World Wide Web,* pp. 519-528, 2003.

[8] X. Ding, B. Liu, and P.S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining," *Proc. Int'l Conf. Web Search and Web Data Mining,* 2008.

[9] Z. Dong and Q. Dong, *Hownet and the Computation of Meaning.* World Scientific Publishing Co., Inc., 2006.

[10] O. Feiguina and G. Lapalme, "Query-Based Summarization of Customer Reviews," *Proc. 20th Conf. Canadian Soc. for Computational Studies of Intelligence on Advances in Artificial Intelligence,* pp. 452-463, 2007.

[11] P. Fragkou, V. Petridis, and A. Kehagias, "A Dynamic Programming Algorithm for Liner Text Segmentation," *J. Intelligent Information System,* vol. 23, no. 2, pp. 179-197, 2004.

[12] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic Recognition of Multi-Word Terms: The C-Value/NC-Value Method," *Int'l J. Digital Libraries,* vol. 3, pp. 115-130, 2000.

[13] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining Customer Opinions from Free Text," *Proc. Sixth Int'l Symp. Intelligent Data Analysis,* pp. 121-132, 2005.

[14] V. Hatzivassiloglou and K. McKeown, "Predicting the Semantic Orientation of Adjectives," *Proc. 35th Ann. Meeting of the Assoc. for Computational Linguistics and Eighth Conf. European Chapter of the Assoc. for Computational Linguistics,* 1997.

[15] M.A. Hearst, "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages," *Computational Linguistics,* vol. 23, no. 1, pp. 33-64, 1997.

[16] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," *Proc. 19th Nat'l Conf. Artificial Intelligence,* 2004.

[17] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 168-177, 2004.

[18] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proc. Int'l Conf. World Wide Web,* pp. 342-351, 2005.

[19] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," *Proc. AAAI-98 Workshop Learning for Text Categorization,* 1998.

[20] E. Milios, Y. Zhang, B. He, and L. Dong, "Automatic Term Extraction and Document Similarity in Special Text Corpora," *Proc. Sixth Conf. Pacific Assoc. for Computational Linguistics,* pp. 275-284, 2003.

[21] S.-M. Kim and E. Hovy, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," *Proc. Workshop Sentiment and Subjectivity in Text,* 2006.

[22] N. Kobayashi, K. Inui, and Y. Matsumoto, "Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining," *Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* pp. 1065-1074, 2007.

[23] M. Krauthammer and G. Nenadic, "Term Identification in the Biomedical Literature," *J. Biomedical Informatics,* vol. 37, no. 6, pp. 512-526, 2004.

[24] B. Pang and L. Lee, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," *Proc. Ann. Meeting on Assoc. for Computational Linguistics,* pp. 115-124, 2005.

[25] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* vol. 2, nos. 1/2, pp. 1-135, 2008.

[26] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing,* 2002.

[27] L. Pevzner and M.A. Hearst, "A Critique and Improvement of an Evaluation Metric for Text Segmentation," *Computational Linguistics,* vol. 28, no. 1, pp. 19-35, 2002.

[28] A.M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," *Proc. Conf. Empirical Methods in Natural Language Processing,* 2005.

[29] G. Qiu, B. Liu, J. Bu, and C. Chen, "Expanding Domain Sentiment Lexicon through Double Propagation," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 1199-120, 2009.

[30] J.C. Reynar, "An Automatic Method of Finding Topic Boundaries," *Proc. 32nd Ann. Meeting Assoc. for Computational Linguistics,* pp. 331-333, 1994.

[31] E. Riloff and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," *Proc. 16th Nat'l Conf. Artificial Intelligence* 1999.

[32] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature Subsumption for Opinion Analysis," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 440-448, 2006.

[33] E. Riloff, J. Wiebe, and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping," *Proc. Seventh Conf. Natural Language Learning at HLT-NAACL,* 2003.

[34] B. Snyder and R. Barzilay, "Multiple Aspect Ranking Using the Good Grief Algorithm," *Proc. Human Language Technology Conf. North Am. Chapter Assoc. of Computational Linguistics,* pp. 300-307, 2007.

[35] V. Stoyanov and C. Cardie, "Topic Identification for Fine-Grained Opinion Analysis," *Proc. 22nd Int'l Conf. Computational Linguistics,* 2008.

[36] M. Thomas, B. pang, and L. Lee, "Get Out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 327-335, 2006.

[37] I. Titov and R. McDonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization," *Proc. Assoc. for Computational Linguistics,* pp. 308-316, 2008.

[38] X. Wan, "Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 553-561, 2008.

[39] B. Wang and H. Wang, "Bootstrapping Both Product Properties and Opinion Words from Chinese Reviews with Cross-Training," *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence,* pp. 259-262, 2007.

[40] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *Proc. 33rd Ann. Meeting Assoc. for Computational Linguistics,* pp. 189-196, 1995.

[41] T. Zagibalov and J. Carroll, "Unsupervised Classification of Sentiment and Objectivity in Chinese Text," *Proc. Third Int'l Joint Conf. Natural Language Processing,* pp. 304-311, 2008.

[42] C. Zhang, K. Wang, M. Zhu, T. Xiao, and J. Zhu, "NEUOM: Identifying Opinionated Sentences in Chinese and English Text," *Proc. Seventh NTCIR Workshop Meeting Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access,* pp. 314-317, 2008.

[43] J. Zhu, H. Wang, B.K. Tsou, and M. Zhu, "Multi-Aspect Opinion Polling from Textual Reviews," *Proc. ACM Conf. Information and Knowledge Management,* pp. 1799-1802, 2009.

[44] L. Zhuang, F. Jing, and X.Y. Zhu, "Movie Review Mining and Summarization," *Proc. 15th ACM Int'l Conf. Information and Knowledge Management,* pp. 43-50, 2006.

Jingbo Zhu received the PhD degree in computer science from Northeastern University, Shenyang, China, in 1999. He has been with the Institute of Computer Software and Theory at Northeastern University since 1999. Currently, he is a full professor in the Department of Computer Science and is in charge of research activities within the Natural Language Processing Laboratory. He was a visiting scholar at ISI, University of Southern California, Los Angeles, from 2006 to 2007. He has published more than 100 papers and holds four US patents. His current research interests include syntactic parsing, machine translation, sentiment analysis, and machine learning for natural language processing. He is a member of the IEEE.

Huizhen Wang received the PhD degree in computer science from Northeastern University, Shenyang, China, in 2008. She has been with the Institute of Computer Software and Theory at Northeastern University since 2008. Currently, she is a lecturer in the Department of Computer Science. She has published more than 30 papers. Her current research interests include knowledge engineering, opinion mining, and machine learning for natural language processing.

Muhua Zhu received the MA degree in computer science from Northeastern University, Shenyang, China, in 2006, where he is currently working toward the PhD degree in the Department of Computer Science. He has published more than 10 papers. His current research interests include syntactic parsing and machine learning for natural language processing.

Benjamin K. Tsou received the MA degree from Harvard University, Cambridge, Massachusetts, and the PhD degree from the University of California, Berkeley. He is a member of the Royal Academy of Overseas Sciences of Belgium and the chair professor of Linguistics and Asian Languages, as well as the director of the Language Information Sciences Research Center of the City University of Hong Kong. Since 1995, he has developed and cultivated the largest (350 million characters, 1.5 million word types by 2008) synchronous corpus of Chinese LIVAC (http://www.livac.org), which makes unique provisions for monitoring linguistic and related trends for application in NLP. He has authored several books and monographs, and more than 100 articles, mostly on computational linguistics and Chinese linguistics. He has served on the editorial boards of *Natural Language Processing* (Japan), the *International Journal of Computational Linguistics and Chinese Language Processing* (IJCLCLP) (Taipei), the *International Journal of Computer Processing of Oriental Languages* (Singapore), and the monograph series on natural language processing from John Benjamins (Amsterdam), among others. His current research interests include sentiment analysis, computational lexicography and lexicology, and resource developing and natural language processing. He is the founding president of the Asian Federation of Natural Language Processing (AFNLP) and was the chairman of SIGHAN, ACL. He also serves on the Executive Board of the Chinese Information Processing Society of China. He is a member of the IEEE.

Matthew Ma received the BS degree in electrical engineering from Tsinghua University, Beijing, China, the MS degree in electrical engineering from the State University of New York at Buffalo, and the PhD degree in electrical and computer engineering from Northeastern University, Boston. He is currently with Scientific Works, Princeton Junction, New Jersey, as a chief scientist. Prior to that, he had 11 years tenure as a senior scientist at Panasonic R&D Company of America, focusing on mobile and imaging research, and two other firms focusing on patent research and strategy. He is the inventor of 14 granted US patents and is the author of more than 30 conference and journal publications. He has published two books in the field of pattern recognition: *Personalization and Recommender Systems* (World Scientific, 2008) and *Mobile Multimedia Processing: Fundamentals, Methods and Applications* (Springer, 2010). He also authored *Fundamentals of Patenting and Licensing for Scientists and Engineers* (World Scientific, 2009). He is an associate editor of the *International Journal of Pattern Recognition and Artificial Intelligence* (IJPRAI) and a US patent agent admitted to the US Patent Trademark & Patent Office. He has been an affiliated professor at Northeastern University, Shenyang, China, since 2001. His primary research interests include patent research, image analysis, pattern recognition, and natural language processing. He is a senior member of the IEEE.