Multi-aspects Rating Prediction Using Aspect Words and Sentences

Takuto Nakamuta **Kazutaka Shimada**
Department of Artificial Intelligence
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka 820-8502 Japan
shimada@pluto.ai.kyutech.ac.jp

Abstract

In this paper we propose a method for a rating prediction task. Each review consists of several ratings for a product, namely aspects. To predict the ratings of the aspects, we utilize not only aspect words, but also aspect sentences. First, our method detects aspect sentences by using a machine learning technique. Then, it incorporates words extracted from aspect sentences with aspect word features. For estimating aspect likelihood of each word, we utilize the variance of words among aspects. Finally, it generates classifiers for each aspect by using the extracted features based on the aspect likelihood. Experimental result shows the effectiveness of features from aspect sentences.

1 Introduction

As the World Wide Web rapidly grows, a huge number of online documents are easily accessible on the Web. Finding information relevant to user needs has become increasingly important. The most important information on the Web is usually contained in the text. We obtain a huge number of review documents that include user's opinions for products. Buying products, users usually survey the product reviews. More precise and effective methods for evaluating the products are useful for users. Many researchers have recently studied extraction and classification of opinions, namely sentiment analysis (Pang and Lee, 2008).

For sentiment analysis, one of the most primitive studies is to classify a document into two classes; positive and negative opinions (Pang et al., 2002; Turney, 2002). One simple extension of p/n classification is a rating prediction task. It is a finer-grained task, as compared with the p/n classification. Several researchers have challenged rating prediction tasks in reviews (Goldberg and Zhu, 2006; Li et al., 2011; Okanohara and Tsujii, 2005; Pang and Lee, 2005). They are called "seeing stars." These tasks handled an overall rating in the prediction. However, each review contains many descriptions about several aspects of a product. For example, they are "performance", "user-friendliness" and "portability" for laptop PCs and "script", "casting" and "music" for movies. Since reviewers judge not only the overall polarity for a product but also details for it, predicting stars of several aspects in a review is also one of the most important tasks in sentiment analysis, instead of a single overall rating. There are several studies to predict some stars in a review, namely "seeing several stars" or "aspect ratings" (Gupta et al., 2010; Pappas and Popescu-Belis, 2014; Shimada and Endo, 2008; Snyder and Barzilay, 2007).

In this paper we propose a method for a rating prediction task with some aspects. In other words, we focus on multi-scale and multi-aspects rating prediction for reviews. We handle video game reviews with seven aspects and zero to five stars. Here we also focus on feature extraction for the prediction. The most common approach is usually based on feature extraction from all sentences in each review. However, all sentences in a review do not always contribute to the prediction of a specific aspect in the review. In other words, the methods handling a review globally are not always suitable to gener-

29th Pacific Asia Conference on Language, Information and Computation pages 513 - 521
Shanghai, China, October 30 - November 1, 2015

ate a model for rating prediction. In addition, Pang and Lee (2004) mentioned that classifying sentences in documents into subjective or objective was effective for p/n classification. In a similar way, for the aspect rating tasks, aspect identification of each sentence and use of aspect sentences for feature extraction might contribute to the improvement for rating prediction. Therefore, the proposed method identifies the aspect of each sentence in each review first. Then, it extracts features for prediction models of seven aspects from all sentences and aspect sentences, on the basis of the variance of words. Finally, it generates prediction models based on Support Vector Regression (SVR) for seven aspects.

2 Related work

Snyder and Barzilay (2007) have proposed a method for multiple aspect ranking using the good grief algorithm. The method utilized the dependencies among aspect ratings to improve the accuracy. Gupta et al. (2010) also have reported methods for rating prediction. They discussed several features and methods for a restaurant review task. They also modified the method based on rating predictors and different predictors for joint assignment of ratings. These methods did not always focus on aspects of each word in reviews.

Shimada and Endo (2008) have proposed a method based on word variance for seeing several stars. They focused on aspect likelihood of each word. The basic idea of our method in this paper is also based on the variance of words in each aspect. However, they computed the variance from all sentences in reviews. On the other hand, our method also focuses on aspect sentences for the computation of the word variance. Pappas and Popescu-Belis (2014) have proposed a method using multiple-instance learning for aspect rating prediction. Their method estimated the weight of each sentences for the prediction. The weights led to the explanation of each aspect. They estimated the aspect weights of each sentence directly in their model. On the other hand, our method identifies the aspect of each sentence by using a machine learning method separately.

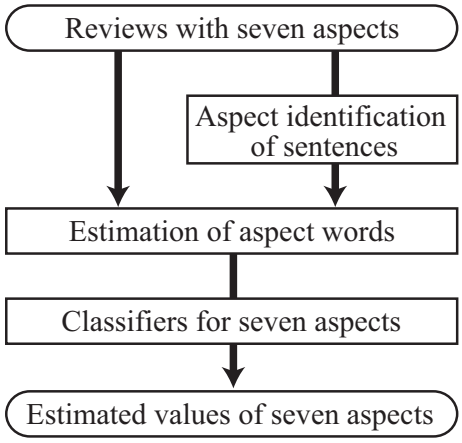

Figure 1: The outline of our method.

3 The proposed method

In this section, we explain the proposed method. Figure 1 shows the outline of our method. It consists of two parts; aspect identification of sentences and estimation of aspect likelihood of words. First, our method identifies the aspects of each sentence in reviews. Then, it estimates aspect likelihood of each word for each aspect, namely aspect words and the weight for each aspect, from aspect sentences and all sentences in reviews. Finally, it generates classifiers for each aspect by using the extracted features based on the aspect likelihood.

3.1 Target data

There are many review documents of various products on the Web. In this paper we handle review documents about video games. Figure 2 shows an example of a review document. The review documents consist of evaluation criteria, their ratings, positive opinions (pros text), negative opinions (cons text) and comments (free text) for a video game. The number of aspects, namely evaluation criteria, is seven: “Originality (o)”, “Graphics (g)”, “Music (m)”, “Addiction (a)”, “Satisfaction (s)”, “Comfort (c)”, and “Difficulty (d)”. The range of the ratings, namely stars, is zero to five points.

We extract review documents from a Web site[1]. The site establishes a guideline for contributions of reviews and the reviews are checked on the basis of the guideline. As a result, the reviews unfitting for

[1] http://ndsmk2.net

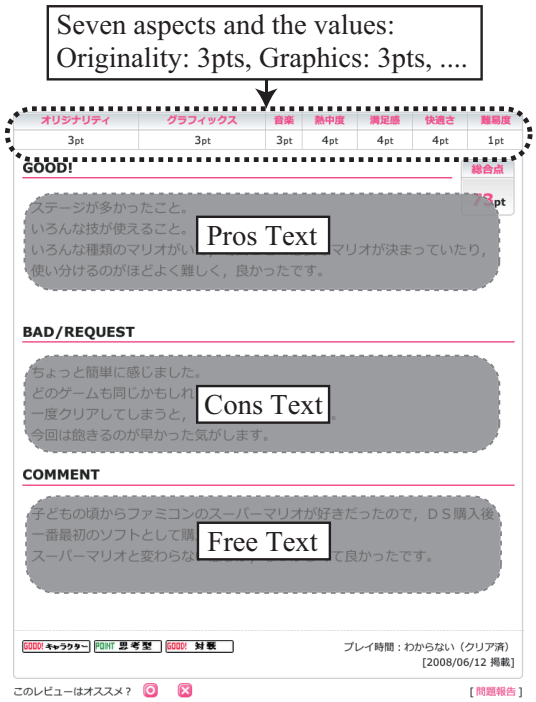

Figure 2: An example of a review document.

the guideline are rejected. Therefore the documents on the site are good quality reviews.

3.2 Aspect identification

First, we identify the aspects of sentences in reviews. For the purpose, we need to construct a aspect-sentence corpus. One annotator detects an evaluative expression from reviews. Then, the annotator selects not only sentences but also short phrases as the evaluative expression. Next, the annotator gives the annotation tags to the detected expression. The annotation tag consists of the polarity and the aspect. Some sentences contain multiple aspect tags. Figure 3 shows examples of the annotation.

We apply a simple machine learning approach with BOW features for the identification process. We employ Support Vector Machine (SVM) as the machine learning approach (Vapnik, 1995). We use nouns, adjectives and adverbs as features for SVM. The feature vector is as follows:

$$f = \{w_1^a, w_2^a, ..., w_{n_a}^a, w_1^c, ...w_{n_c}^c, ..., w_1^s, ..w_{n_s}^s\}$$

where w^x denotes a word w in an aspect x, and $x \in \{a, c, d, g, m, o, s\}$ (See Section 3.1). n_x denotes the number of words appearing in an aspect x.

<m p>The music is incredibly powerful sound!</>
Positive (p) about the aspect “Music (m)”

<s n>It lacks a feeling of accomplishment after finishing.</>
Negative (n) about the aspect “Satisfaction (s)”

<g,s p,p>Since the graphics was beautiful, we got the satisfaction from just watching them.</>
Combined tags are acceptable:
Positive (p) about the aspects “Graphics (g)” and “Satisfaction (s)”

Figure 3: Examples of aspect annotation of sentences.

The vector value of a word is computed as follows:

$$val(asp_i, w_j) = \frac{num_{ij}}{sent(asp_i)} \quad (1)$$

where num_{ij} and $sent(asp_i)$ denote the frequency of a word w_j in an aspect asp_i and the number of sentences belonging to an aspect asp_i, respectively. This is a normalization process because the numbers of sentences belonging to each aspect are non-uniform. We generate seven classifiers for seven aspects using the features and values; the classifier for the aspect “Addiction (a)” or not, the classifier for the aspect “Comfort (c)” or not, and so on. Figure 4 shows the aspect identification process[2]. We use the SVM^{light} package[3] with all parameters set to their default values (Joachims, 1998).

3.3 Rating prediction

Removing non-informative text from training data leads to the improvement of the accuracy (Fang et al., 2010). In this task, a word does not always contribute to all aspects. A word usually relates to one or two aspects. Therefore, estimating a relation between a word and each aspect is the most important task for the rating prediction. It improves the performance.

We introduce a variance-based feature selection proposed by (Shimada and Endo, 2008) into this process. They obtained small improvement in terms of an error rate by using the variance-based feature selection. The basic idea is to extract words appearing frequently with the same point (stars) regarding

[2]Note that the method does not estimate the polarity, namely positive or negative, in this process.

[3]http://svmlight.joachims.org

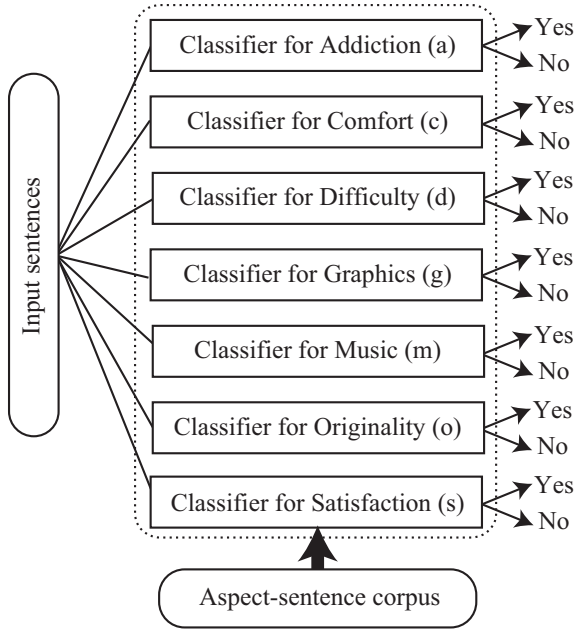

Figure 4: The sentence-aspect identification.

an evaluation criterion (aspect). It is computed as follows:

$$var(w_{a_j}) = \frac{1}{m} \sum_{i=0, w \in r_i}^{n} (real(r_i, a_j) - ave(w_{a_j}))^2 \quad (2)$$

where a_j is an aspect. m and n are the document frequency (df) of a word w and the number of documents respectively. $real(r_i, a_j)$ and $ave(w_{a_j})$ are the actual rating of a_j in r_i and the average score of w for a_j. We use w of which the var is a threshold or less.

We apply the variance-based feature selection to aspect sentences extracted in Section 3.2 and all sentences in pros and cons text areas[4]. We use MeCab for the morphological analysis[5]. We select words belonging to "noun", "adjective" and "adverb". Finally, we extract words as features on the basis of the word frequency ($freq$) and the value var. In addition, we distinguish words in the pros text areas and the cons text areas. In other words, for a word w_i, a word in the pros text areas is w_i^p and a word in the cons text areas is w_i^c. Besides, we distinguish words from all sentences ($w_i^{x^{al}}$) and aspect-sentences ($w_j^{x^{ap}}$). i and j are the numbers of words from all sentences and aspect-sentences, respectively. A vector of an aspect y for a review x is as follows:

$$r_{xy} = \{w_1^{p^{al}}, w_2^{p^{al}}, ..., w_i^{p^{al}}, w_1^{c^{al}}, w_2^{c^{al}}, ..., w_i^{c^{al}}, w_1^{p^{ap}}, w_2^{p^{ap}}, ..., w_j^{p^{ap}}, w_1^{c^{ap}}, w_2^{c^{ap}}, ..., w_j^{c^{ap}}\}$$

We apply the vector into a machine learning approach. In this paper, we employ a linear support vector regression (SVR). This is one of straightforward methods for this task. Related studies also used SVR for the rating inference task (Okanohara and Tsujii, 2005; Pang and Lee, 2005; Shimada and Endo, 2008). We generate seven classifiers for seven aspects using the selected features. We also use the SVM^{light} for SVR.

4 Experiment

In this section, we describe two experiments about the aspect identification of sentences and the rating prediction. For the rating prediction, we evaluate the effectiveness of the aspect-sentences.

4.1 Aspect identification

The annotated corpus for the aspect identification consisted of 4719 sentences. Table 1 shows the distribution of aspects[6]. The table shows that there were large differences among aspects. Machine learning with unbalanced data usually leads to generation of a wrong classifier. Therefore, we adjusted the number of sentences in the training data (use_s) for each classifier by using the following equation.

$$use_s(asp_i, asp_j) = real_s(asp_j) \times \frac{real_s(asp_i)}{all_s - real_s(asp_i)} \quad (3)$$

where asp_i and asp_j denote the target aspect and the others, respectively. $real_s(asp_j)$ denotes the number of sentences of an aspect asp_j and all_s is the number of sentences in the corpus, 4719 in this experiment. The instance about Addiction (a) is shown in Table 2. Since the number of sentences in the Addiction (a), asp_i, was 429, the sum of the others was 427.

We evaluated our method with 10-fold cross validation. The criteria are the precision, recall and F-value. Table 3 shows the experimental result. The

[4]We ignore sentences in the free text area in Fig. 2.

[5]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[6]Note that more than half of sentences in the corpus contained two or three aspects.

Aspect	# of sentences
Addiction (a)	429
Comfort (c)	354
Difficulty (d)	353
Graphics (g)	230
Music (m)	258
Originality (o)	2339
Satisfaction (s)	2252

Table 1: The aspects and the number of sentences.

Aspect	Original	Training	
Addiction (a)	429	429	
Comfort (c)	354	26	427
Difficulty (d)	353	26	
Graphics (g)	230	17	
Music (m)	258	19	
Originality (o)	2339	173	
Satisfaction (s)	2252	166	

Table 2: Downsized and adjusted training data for Addiction (a)

aspects "Originality" and "Satisfaction" obtained comparatively higher accuracy rates because they consisted of sufficient training data. Sentences of the aspect "Graphics" tended to contain direct expressions related to graphics, such as "beautiful." In addition, they were usually simple sentences; "The graphics are ... ." The aspect identification about the aspects "Addiction", "Comfort" and "Difficulty" were difficult tasks. In comparison with the aspect "Graphics", sentences of these aspects did not always contain direct expressions; e.g., "I play this game every day" for "Addiction", "There are many situations about pressing A when I need to push B" for "Comfort", and "The enemy in the water area is too clever" for "Difficulty." This was one reason that the recall rates of them were extremely low, as compared with others. It is difficult to identify these aspects correctly, especially with a small dataset.

4.2 Rating prediction

Next, we evaluated our method for the rating prediction. We prepared three different sizes of training data; (ds1) 933 reviews about 7 games, (ds2) 2629 reviews about 37 games and (ds3) 3464 reviews about 44 games. They were balanced data sets. In other word, each data set contained reviews about products with high and low scores uniformly. These data sets did not contain any reviews that were used in the aspect identification of sentences in Section 4.1. For the determination of the thresholds about the aspect likelihood var and the word frequency ($freq$) in Section 3.3, we also prepared the development data set consisting of 76 reviews. If we set high thresholds for them, we might obtain features with high confidence about each aspect. However, too thigh thresholds usually generate a zero-vector, which does not contain any features. We estimated these thresholds, which did not generate a zero-vector, from the development data. In this experiment, var and $freq$ for all sentences were less than 1.5 and more then 3, and var and $freq$ for aspect-sentences were less than 0.5 and more than 4, respectively.

Aspect	Precision	Recall	F-value
Addiction (a)	0.941	0.186	0.310
Comfort (c)	0.772	0.249	0.377
Difficulty (d)	0.738	0.272	0.398
Graphics (g)	0.890	0.630	0.738
Music (m)	0.404	0.353	0.377
Originality (o)	0.805	0.559	0.660
Satisfaction (s)	0.746	0.562	0.641
Average	0.756	0.402	0.525

Table 3: The experimental result of aspect identification.

We evaluated our method with the leave-one-out cross-validation for the three data sets. The criterion for the evaluation was the mean squared error (MSE).

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(out(d_{ij}) - real(d_{ij}))^2 \qquad (4)$$

where i and j denote a review and an aspect in the review respectively. out and $real$ are the output of a method and the actual rating in a review respectively. We converted the outputs of the SVR into integral value with half adjust because it was continuous. The MSE is one of important criteria for the rating inference task because not all mistakes of estimation with the methods are equal. For example, assume that the actual rating of a criterion is 4.

Aspect	data (ds1)		data (ds2)		data (ds2)	
	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
Addiction (a)	1.146	**1.047**	1.203	**1.054**	1.288	**1.068**
Comfort (c)	0.887	**0.881**	0.975	**0.944**	0.980	**0.901**
Difficulty (d)	**0.855**	0.856	0.888	**0.872**	**0.864**	0.866
Graphics (g)	0.704	**0.674**	0.693	**0.644**	0.711	**0.677**
Music (m)	0.665	**0.654**	0.719	**0.666**	0.715	**0.671**
Originality (o)	**0.770**	0.772	**0.757**	0.766	0.789	**0.759**
Satisfaction (s)	1.296	**1.110**	1.210	**1.036**	1.266	**1.055**
Average	0.903	**0.856**	0.921	**0.854**	0.944	**0.857**

Table 4: The experimental result of the rating prediction.

In this situation, the mistake of estimating it as 3 is better than the mistake of estimating it as 1.

We compared our method[7] with a baseline. The baseline did not use any aspect-sentence information. In other words, it was based on (Shimada and Endo, 2008). Table 4 shows the experimental result. Our method outperformed the baseline for all data sets. The improvements were 0.047 (approximately 5% on the error rate) for the data (ds1), 0.066 (approximately 7% on the error rate) for the data (ds2) and 0.087 (approximately 9% on the error rate) for the data (ds3). For the data (ds2) and (ds3), our method yielded significant differences at $p < 0.05$ by t-test. The results show the effectiveness of the aspect identification of sentences and the feature extraction based on the aspect-sentences. In addition, the MSE values on the proposed method were stable although those on the baseline decreased when the size of the data set was changed. This result show the proposed method is robust in the case that noise in training data increases.

4.3 Discussion

A review does not always consist of sentences related to all aspects. Reviews often do not contain any sentences related to an aspect. Gupta et al. (2010) reported that only 62% of user given ratings have supporting text for ratings of the aspects in their review data. In (Shimada and Endo, 2008), it was approximately 75% in their dataset, which was similar to our dataset. Therefore, we computed the content rate of aspect-sentences in each data set. The rate is computed by

$$CR = \frac{NumAspRev}{NumRev} \quad (5)$$

where $NumAspRev$ denotes the number of reviews which contain identified aspect-sentences. $NumRev$ is the number of reviews about an aspect in the data set.

We computed the CR values for the three data sets and the development data. Table 5 shows the CR values of all aspects on each data set. The CR values on the development data was a kind of oracle situation because the sentences in the data were annotated by human. From the CR on the development in Table 5, approximately 30% of reviews in our data set were missing the textual support for some aspects in the reviews. This is one reason that the MSE values in Section 4.2 were not sufficient. In other words, owing to lack of textual information, the aspect rating prediction is essentially a difficult task.

The CR values of the aspects "Addiction", "Comfort" and "Difficulty" on the three test data set were lower than the development data. The accuracy of the aspect identification in Table 3 shows a similar trend. On the other hand, the CR of the aspect "Music" was too high, as compared with the development data. This was caused by the low precision rate of the aspect identification (also see Table 3). To improve the accuracy of the aspect identification leads to the improvement of the rating prediction. The improvement of these recall and precision rates for these aspects is one of the important tasks.

As you can see from Table 5, the rating prediction

[7]Note that the method used the aspect-sentences identified automatically in the previous section. They were not oracle data.

Aspect	development	data (ds1)	data (ds2)	data (ds3)
Addiction (a)	0.750	0.330	0.340	0.337
Comfort (c)	0.934	0.229	0.307	0.287
Difficulty (d)	0.631	0.227	0.231	0.232
Graphics (g)	0.408	0.410	0.426	0.424
Music (m)	0.237	0.478	0.477	0.479
Originality (o)	0.961	0.927	0.961	0.968
Satisfaction (s)	0.961	0.912	0.954	0.958
Average	0.697	0.502	0.528	0.526

Table 5: The content rate of aspect-sentences.

in the proposed method used only approximately 50% of the identified aspect-utterances. Moreover, 25% of sentences in the aspect identification were wrong (see the average precision rate in Table 3). Despite the fact that the input data of the rating prediction contained many mistakes, the proposed method with aspect-sentences outperformed the baseline without aspect-sentences. The result shows that the aspect-sentences are essentially effective to predict aspect ratings even if they contain misrecognized data. If the accuracy of the aspect identification is improved, the accuracy of the rating prediction is also improved. Therefore, the improvement of the aspect identification is the most important future work. The identification task in our study is a multi-label classification problem. Applying multi-label learning such as (Zhang and Zhou, 2007) to the task is one of the most interesting approaches although we used a binary classifier based on SVMs. Another problem in the identification task was the unbalance data. As we mentioned in Section 4.1, we handled this problem by adjusting the number of sentences in the training data. Under such circumstances, Complement Naive Bayes (CNB) (Rennie et al., 2003) is often effective. Applying this method to the task is interesting. Besides, we applied a classification method in the identification task. The recall rate was not sufficient. An extraction approach based on bootstrapping (Etzioni et al., 2004; Riloff and Jones, 1999), which uses the extracted aspect-sentences as seeds, is also an interesting approach to obtain more aspect sentences in the data.

In this experiment, we used SVR to estimate the ratings in a document. The SVR is often utilized in rating inference tasks. However, Pang and Lee (2005) have proposed a method based on a metric labeling formulation for a rating inference problem. The results of these studies denote that SVR is not always the best classifier for this task. Koppel and Schler (2006) have discussed a problem of use of regression for multi-class classification tasks and proposed a method based on optimal stacks of binary classifiers. Tsutsumi et al. (2007) have proposed a method based on the combination of several methods for sentiment analysis. We need to consider other methods for the improvement of the accuracy.

We estimated aspect likelihood based on a variance of each word. Kobayashi et al. (2004) have proposed a method to extract attribute-value pairs from reviews. The attributes relate to aspects in our work. Wilson et al. (2004) have proposed a method to classify the strength of opinions. Sentiment word dictionaries with aspects and strength are useful for the rating prediction. Besides, Kobayashi et al. (2005) have expanded their work with an anaphora resolution technique. To identify the aspect of a sentence more correctly, context information in reviews is also important.

In this paper, the aspects for the rating prediction are given. Yu et al. (2011) have proposed an aspect ranking method for reviews. They identified important product aspects automatically from reviews. Aspect mining is also interesting future work.

5 Conclusion

In this paper we proposed a multi-scale and multi-aspects rating prediction method based on aspect-sentences. The target reviews contained seven aspects with six rating points. Despite the fact that the

input data of the rating prediction contained many mistakes, namely lack of 50% and misrecognition of 25%, the proposed method with aspect-sentences outperformed the baseline without aspect-sentences. The experimental results show the effectiveness of the aspect identification of sentences in reviews for the rating prediction. Therefore, the improvement of the aspect identification of sentences is the most important future work.

In this paper, we dealt with only predicting ratings in reviews. However, estimating relations between aspects and words is beneficial for many sentiment analysis tasks. Yu et al. (2011) reported that the extracted aspects improved the performance of a document-level sentiment classification. Applying the result and knowledge from the rating prediction in this paper to other tasks, such as summarization (Gerani et al., 2014; Shimada et al., 2011), is also interesting future work.

References

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall (preliminary results). In *Proceedings of the 13th international conference on World Wide Web (WWW2004)*, pages 100–110.

Ji Fang, Bob Price, and Lotti Price. 2010. Pruning non-informative text through non-expert annotations to improve sentiment classification. In *Coling 2010 Workshop: The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613.

Andrew B. Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52.

Narendra Gupta, Giuseppe Di Fabbrizio, and Patrick Haffner. 2010. Capturing the stars: Predicting ratings for service and product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 36–43.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the First International Joint Conference on Natural Language Processing*, IJCNLP'04, pages 596–605.

Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2005. Opinion extraction using a learning-based anaphora resolution technique. In *In The Second International Joint Conference on Natural Language Processing (IJCNLP)*, pages 175–180.

Moshe Koppel and Jonathan Schler. 2006. The importance of neutral examples in learning sentiment. *Computational Intelligence*, 22(2):100–109.

Fangtao Li, Nathan Liu, Hongwei Jin, Kai Zhao, Qiang Yang, and Xiaoyan Zhu. 2011. Incorporating reviewer and product information for review rating prediction. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 1820–1825.

Daisuke Okanohara and Jun'ichi Tsujii. 2005. Assigning polarity scores to reviews using machine learning techniques. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, pages 314–325.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and TrendsR in Information Retrieval*, 2.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pages 455–466.

Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *In Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 616–623.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceeding of AAAI 99*, pages 474–479.

Kazutaka Shimada and Tsutomu Endo. 2008. Seeing several stars: A rating inference task for a document containing several evaluation criteria. In *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008*, pages 1006–1014.

Kazutaka Shimada, Ryosuke Tadano, and Tsutomu Endo. 2011. Multi-aspects review summarization with objective information. *Procedia - Social and Behavioral Sciences: Computational Linguistics and Related Fields*, 27:140–149.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pages 300–307.

Kimitaka Tsutsumi, Kazutaka Shimada, and Tsutomu Endo. 2007. Movie review classification based on a multiple classifier. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 481–488.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 761–767.

Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect ranking: Identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1496–1505.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048.