SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 22 – 31

Procedia
Social and Behavioral Sciences

Pacific Association For Computational Linguistics (PACLING 2011)

Aspect identification of sentiment sentences using a clustering algorithm

Masashi Hadano[a], Kazutaka Shimada[a*], Tsutomu Endo[a]

[a]*Department of Artificial Intelligence, Kyushu Institute of Technology*
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Abstract

Reviews contain aspect information of a product, such as “image quality” and “usability" of a camera. In this paper, we propose an aspect identification method for sentiment sentences in review documents. Machine learning methods usually require a large amount of training data for generating a classifier with high accuracy. However, preparing training data by hand is costly. To solve this problem, we apply a clustering approach to the aspect identification method. Our system acquires new training data from non-tagged data by using the clustering approach. As compared with a baseline method, which does not use the acquisition approach, our method obtained high accuracy.

Keywords: Sentiment analysis; Aspect identification; Clustering

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 of products. When buying a product, users usually survey reviews of the product. More precise and effective methods for evaluating the products are useful for users. To

* Corresponding author. Tel.:+81-948-29-7636.
E-mail address: shimada@pluto.ai.kyutech.ac.jp

doi:10.1016/j.sbspro.2011.10.579

analyze the opinions is one of the hottest topics in natural language processing. Many researchers have recently studied extraction of evaluative expressions and classification of opinions [1, 2, 3, 9].

Here we focus on aspects in opinions about a target product. The aspects denote an attribute or evaluative point of the target, such as “image quality” and “usability” of a camera. To identify the aspect of sentences in reviews is an important task for applications of sentiment analysis. Tadano et al. [7] have proposed a multi-review summarization system focusing on the aspect of each sentence in reviews. Blair-Goldensohn et al. [1] have also reported a sentiment summarizer with aspect information for local service reviews.

In this paper, we propose an aspect identification method for sentiment sentences in review documents. In general, machine learning techniques or statistical approaches are employed for the identification or classification tasks. They usually require a large amount of training data for generating a classifier with high accuracy. However, preparing training data by hand is costly. To solve this problem, we propose a method that acquires new training data from non-tagged data by using a clustering approach. By using this method, we can improve the performance of a classifier with a small training data set. Figure 1 shows the outline of our method. First, it classifies similar sentences into clusters. Then, a user tags the aspect of sentences which are close to the centroid of each cluster. Our method obtains new training data by using the tagged sentences. Finally, we identify the aspect of sentences in test data by using a machine learning method, SVM, with the new training data.

In Section 2, we explain the data set in this paper. In Section 3, we describe the aspect identification process and evaluate our method in Section 4. Finally, we conclude the paper in Section 5.

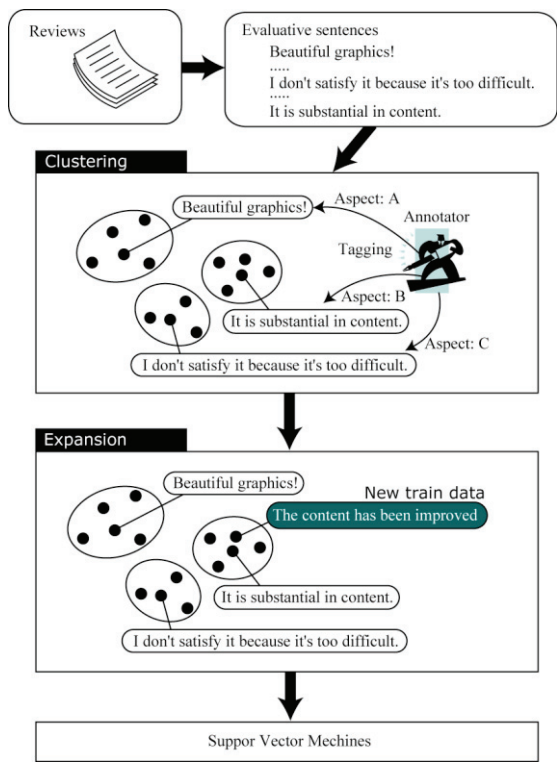

Fig. 1. The outline of our method.

2. Data set

We use game review documents as target data. The review documents were extracted manually from the Web site[1]. Seven evaluative criteria are given to each review, i.e., Originality (o), Graphics (g), Music (m), Addiction (a), Satisfaction (s), Comfort (c), and Difficulty (d). Figure 2 shows an example of the review document. It consists of positive and negative opinions with evaluative criteria.

For generating test data, we use a tool for constructing a sentiment corpus which is proposed by Tadano et al. [6]. The tool needs pre-annotated data for the support of a current annotation process. Therefore two annotators (A1, A2) beforehand annotated the review documents. The annotated data consists of 3,446 expressions by A1 and 1,589 expressions by A2. The rate of which both annotators detected the same expression was 42.7% and the rate of which annotators gave the same tag was 0.456 on kappa value. Although the initial agreement is not sufficient, the annotation tool boosts up the agreement of the annotation process. They reported that the agreement and kappa value were improved to 85.7% and 0.687 by using the tool in [6].

In this paper, one annotator constructs a data set for evaluation by using the annotation tool. Firstly, the annotator detects an evaluative expression from a document. The annotator selects not only sentences but also short phrases as the evaluative expression. Then, the annotator gives the annotation tags to the detected expression. The annotation tag consists of the polarity and the evaluative criteria. The evaluative criteria tag consists of the seven kinds. Several evaluative criteria tags may be given to the same expression. Figure 3 shows an example of the actual annotation. In this paper, we regard the evaluative criteria as aspects.

The annotated data contains a wide distribution of combinations of aspects; e.g., *d* and *s* in Figure 3. Therefore, several combined aspects consist of a few sentences. In this paper, we handle the combinations that possess 10 sentences or more. As a result, the target data consists of 4607 sentences with 20 aspect combinations from 485 reviews. Figure 4 shows the distribution of the combination.

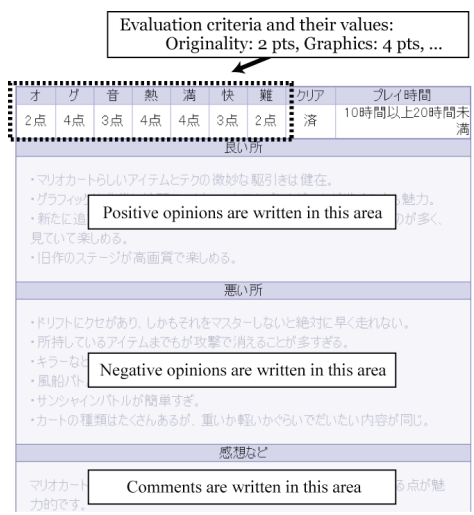

Fig. 2. An example of the review document.

[1] http://ndsmk2.net/

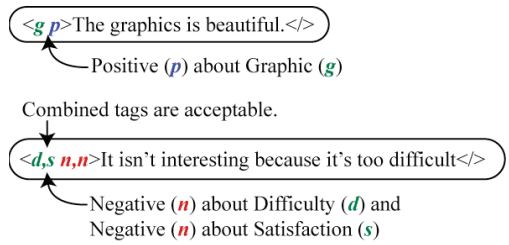

Fig. 3. An example of the annotation.

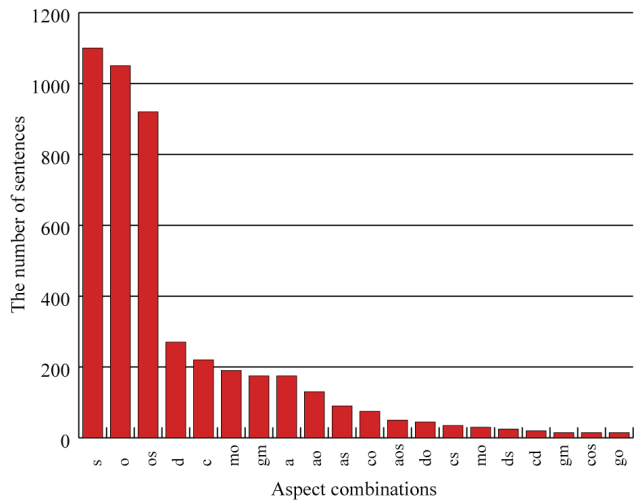

Fig. 4. The distribution.

3. Proposed method

In this section, we explain the proposed method. It consists of four processes: (1) Clustering of sentences, (2) Annotation of initial training data, (3) Acquisition of new training data and (4) Classification with machine learning.

3.1. Clustering

In this paper, we start with an assumption that similar sentences contain the same aspect. Therefore, we organize similar sentences by using a clustering method. For implementation of the clustering, we use Bayon[2] which is a simple and fast hard-clustering tool[3]. The clustering of Bayon is based on the repeated bisection algorithm. The process is as follows:

1. divide initial data into two clusters.
2. detect the cluster which possesses the lowest similarity between the centroid and each element in it.
3. select two elements randomly.
4. classify all elements in the cluster selected in step 2 into two clusters on the basis of a similarity between the elements selected in step 3 and elements in the cluster.
5. swap elements between the clusters if it improves the similarity[4].
6. repeat step 2 - 5.

Figure 5 show the process of the clustering. The feature vector for the clustering consists of content words in evaluative sentences.

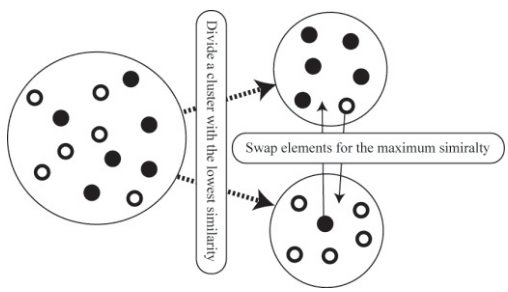

Fig. 5. The process of the clustering.

3.2. Annotation

In this process, an annotator determines the aspect of sentences in clusters generated in the previous section. Each sentence in the clusters contains a similarity value which is calculated by the cosine similarity measure between the vector of the sentence and the vector of the cluster centroid. Our system displays sentences with the maximum similarity in each cluster. Then, an annotator judges one aspect for each sentence. The annotated data is the initial training data. In other words, the number of training data is

[2] http://code.google.com/p/bayon/
[3] In this paper, we used "-l 2.0" and "-idf" options.
[4] Actually, it is the summation of the COS similarity in the cluster

the number of clusters. Figure 6 shows the annotation process. The annotator identifies the aspect of each representative sentence.

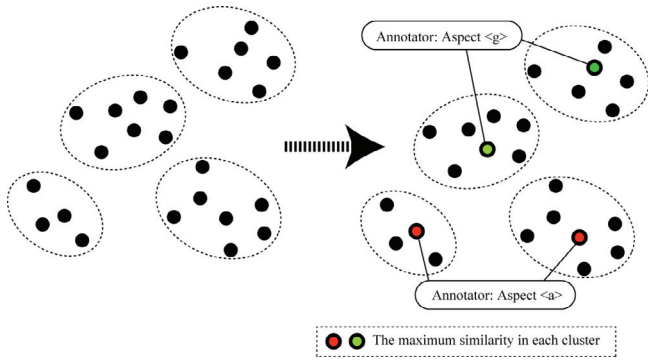

Fig. 6. The process of the annotation.

3.3. Acquisition

In the previous process, namely the annotation process, we obtain the initial training data. It consists of representative sentences with an aspect in each cluster. In other words, the number of annotated sentences is the number of clusters. These annotated aspects contain a high confidence because they are tagged by a human annotator. However, the number of sentences is generally insufficient for generating a classifier with a higher accuracy.

Here there is an assumption that similar sentences contain the same aspect. It denotes that the aspects of sentences belonging to each cluster are equal to the aspects of the representative sentences of each cluster. On the basis of this assumption, we acquire new sentences from each cluster as the new training data. However, the assumption that all sentences belonging to a cluster contain the same aspect is too naive. The clusters often possess sentences with different aspects because the clustering method is imperfect.

In this paper, we focus on the similarity value between a sentence and the centroid in each cluster. In general, sentences with the high similarity value possess high confidence because they are close to a representative sentence. Figure 7 shows the acquisition process. In this case, our system acquires sentences with the high similarity as new training data. These sentences are effective for generating a classifier. However they might not contribute to dramatically improve the accuracy because they are similar to the representative sentences. It is also important to acquire sentences with different expressions. Therefore, we apply the different similarity range to the acquisition process.

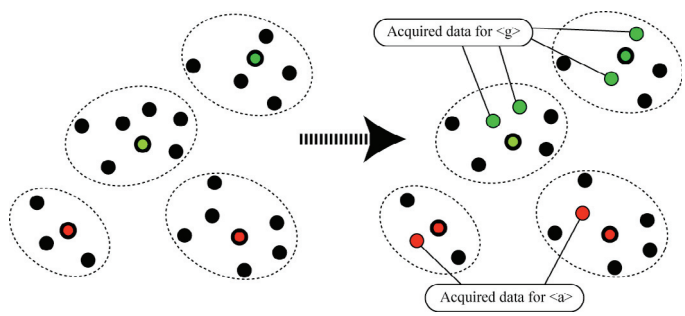

Fig. 7. The process of the acquisition.

3.4. Classification

We use Support Vector Machines (SVMs) as the classifiers. SVMs are a machine learning algorithm that was introduced by [10]. They have been applied to tasks such as face recognition and text classification. An SVM is a binary classifier that finds a maximal margin separating hyperplane between two classes. The hyperplane can be written as:

$$y_i = \vec{w} \cdot \vec{x} + b$$

where x is an arbitrary data point, i.e., feature vectors, w and b are decided by optimization, and y_i in {+1, -1}. The instances that lie closest to the hyperplane are called support vectors. Figure 8 shows an example of the hyperplane. In the figure, the solid line shows the hyperplane. For implementation of the SVMs, we use LIBSVM[5]. The feature set for SVMs consists of content words in each sentence, namely BOW features.

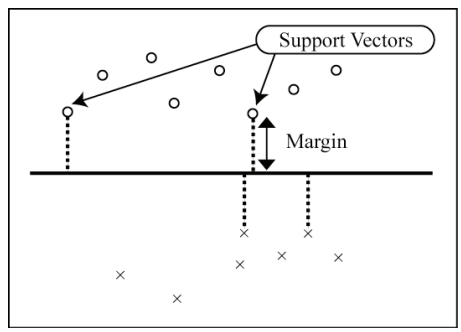

Fig. 8. Support vector machines.

4. Experiment

In this section, we evaluated our method with the data set described in Section 2.

4.1. Settings

The dataset consisted of 4607 sentences annotated manually. We evaluated our method with a partial match accuracy rate although the sentences often contained combined aspects. The accuracy was computed as follows:

$$Accuracy = \frac{\sum_{asp \in A} R(asp)}{\sum_{asp \in A} N(asp)}$$

where A is the 7 basic single aspects, namely Originality (o), Graphics (g), Music (m), Addiction (a), Satisfaction (s), Comfort (c), and Difficulty (d). $N(asp)$ is the number of aspects which are partially contained in the output from our method. $R(asp)$ is the number of correct aspects in $N(asp)$.

We evaluated the data set with 10-folds cross-validation. The average number of clusters was 219 in the cross-validation. For the acquisition process, we compared three types of similarity ranges. In addition, we compared our method with machine learning methods, SVMs and C4.5 [4], using fully annotated data.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4.2. Results

Table 1 shows the experimental result. The baseline denotes an approach that was based on random sampling of initial training data. In the baseline method, an annotator identified the aspect of approximately 200 sentences, which depended on the number of clusters in each validation, extracted randomly from the data set. In other words, it did not use any clustering approaches for the annotation process. The values in "Range" denote the range of the similarity for the acquisition process. "NoExp" was the approach without the acquisition process. In other words, the number of sentences as training data was the number of clusters. "AllSVM" and "AllC4.5" were approaches using fully annotated data. They used approximately 4000 sentences for the learning in each validation step. On the other hand, the average number of training data was 219 sentences for the baseline and the proposed method with "NoExp". In the acquisition process, we obtained approximately 250, 930 and 2200 sentences in the range "0.7-0.9", "0.5-0.7" and "0.3-0.5" respectively as the new training data.

The proposed method outperformed the baseline method. The baseline was based on random sampling for the annotation. The distribution of aspects in our data set was not identically-distributed (See Figure 4). Therefore, it generated a biased training data. As a result, it led to decrease of the accuracy.

Table. 1. Experimental result.

Method	Range	Accuracy
Baseline	-	67.28
Proposed	NoExp	73.80
Proposed	0.7-0.9	**73.97**
Proposed	0.5-0.7	71.30
Proposed	0.3-0.5	67.30
AllSVM	-	80.93
AllC4.5	-	73.86

The proposed method with "0.7-0.9" produced the best performance. The acquisition processes using sentences which were not close to the centroid, namely "0.5-0.7" and "0.3-0.5" were not effective. Moreover, there is small difference between the proposed methods with "0.7-0.9" and "NoExp". Although sentences in the range "0.7-0.9" were often coincident with the aspect of a representative sentence, the contribution to the accuracy was slight. The reason was that the acquisition process just obtained sentences similar to each representative sentence on surface expressions because the clustering process was based on BOW features. It did not lead to improvement of coverage for the training data. To achieve a higher accuracy, we need to discuss a method for the acquisition process that handles semantic information of words.

Although the training data consisted of approximately 200 initial training data by the clustering and annotation processes and 250 sentences extracted in the acquisition process, the accuracy ranked with C4.5 with approximately 4000 sentences as the training data. This result shows the effectiveness of our method with the annotation and acquisition process. On the other hand, the accuracy of "AllSVM" was 80.93. It denotes that the accuracy might be improved to 80.93% if our method obtained more appropriate sentences in the acquisition process. The acquisition process is the most important process in our method.

The annotated data contained a wide distribution of aspects. In this situation, aspects consisting of a large quantity of training data, namely aspect "S" and "O", often generate an undeserved contribution to the whole accuracy. Therefore we computed the standard deviation of the accuracy rates from each

aspect. The value was approximately 4%. This result shows that our method can treat minority aspects correctly.

In the annotation process (Section 3.2), the annotator judged one aspect for each sentence. However, sentences occasionally contained a combined aspect (See Figure 4). Therefore, we should essentially allow the annotation of a combined aspect in this process. On the other hand, we acquired new data by using directly the annotated data in the acquisition process. We think that inheriting the combined aspect to sentences in each cluster is not always appropriate. To solve this problem of combined aspects in the annotation and acquisition is important future work.

In this experiment, we evaluated our method with a partial match accuracy rate because we did not deal with the problem of combined aspects in the proposed method. The evaluation with the complete matching accuracy is an important and challenging task and future work for our method[6].

5. Conclusion

In this paper, we proposed an aspect identification method for sentiment sentences in review documents. To solve this problem of the number of training data, we applied non-tagged data and a clustering approach. Our method classified similar sentences into clusters first. Then, a user tagged the aspect of sentences which are close to the centroid of each cluster. Our method acquired new training data by using the tagged sentences. Finally, we identified the aspect of sentences in test data by using a machine learning method, SVM, with the new training data. The method with the clustering approach outperformed the method without the clustering approach (73.97 vs. 67.28).

The method with approximately 450 sentences as training data was equal to C4.5 with approximately 4000 sentences. This result shows the effectiveness of our method with the clustering and acquisition processes. In addition, our method holds the possibility that it improves the accuracy to approximately 80% because the accuracy of SVMs with full annotated data was 80.93%. To achieve a higher accuracy, we need to discuss the acquisition process.

Our method is in the category of the active learning, which is an algorithm based on interaction with a user [5]. We need to discuss other approaches for the interaction process, namely the annotation and acquisition processes. Titov and McDonald [8] have proposed a joint model of text and aspect ratings for sentiment summaries. Their method was based on the Multi-Grain Latent Dirichlet Allocation model (MG-LDA) and identified the relation between text and aspect without training data. We need to consider unsupervised learning for our task.

Acknowledgment

This work was supported by Kayamori Foundation of Informational Science Advancement.

References

[1] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In WWW Workshop on NLP in the Information Explosion Era (NLPIX).

[6]The complete accuracy of AllSVM was 62.9%. The low accuracy denotes that identifying combined aspects is a tough issue.

[2] Kobayashi, N., Inui, K., and Matsumoto, Y. (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1065–1074.
[3] Pang, B., Lee, L., and Vaithyanathan, S. (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.
[4] Quinlan., J. R. (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers.
[5] Settles, B. (2009). Active learning literature survey. Technical report, Computer Sciences Technical Report 1648, University of Wisconsin, Madison.
[6] Tadano, R., Shimada, K., and Endo, T. (2009). Effective construction and expansion of a sentiment corpus using an existing corpus and evaluative criteria estimation. In Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics (PACLING2009), pages 211–216.
[7] Tadano, R., Shimada, K., and Endo, T. (2010). Multi-aspects review summarization based on identification of important opinions and their similarity. In Proceedings of the 24nd Pacific Asia Conference on Language, Information and Computation (PACLIC24).
[8] Titov, I. and McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In Proceedings of the 46th Meeting of Association for Computational Linguistics (ACL-08), pages 308–316.
[9] Turney, P. D. (2002). Thumbs up? or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 417–424.
[10] Vapnik, V. N. (1999). Statistical Learning Theory. Wiley.