Multi-aspects Review Summarization with Objective Information

Kazutaka Shimada, Ryosuke Tadano and Tsutomu Endo
Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
{shimada, r_tadano, endo}@pluto.ai.kyutech.ac.jp

Abstract

In this paper, we propose a method for multi-aspects review summarization based on evaluative sentence extraction. We handle three features; ratings of aspects, the $tfidf$ value, and the number of mentions with a similar topic. For estimating the number of mentions, we apply a clustering algorithm. By using these features, we generate a more appropriate summary. In this paper, we also focus on objective information of the target product. We integrate the summary from sentiment information in reviews and the objective information extracted from Wikipedia. The experiment results show the effectiveness of our method.

Keywords: sentiment analysis, multi-aspects review summarization, objective information, opinion integration.

1. Introduction

As Web services like CGMs have become widely used, people can easily post reviews for products or services. Although handling these information (evaluative information) has become necessary, there exists too much information on the Web. Therefore, extracting information that users want and summarizing them have been expected recently. Intuitively, we can summarize a review with traditional document summarization methods. For instance, Brandow et al. [2] have summarized a document by extracting sentences with some features such as the presence of signature words and the location in the document. For sentiment summarization, Pang and Lee [9] have extracted all subjective sentences. They suggested that these extracted sentences could be used as summaries. However, a review basically consists of sentiments with various aspects (i.e., “image quality” and “usability” of a camera). Therefore, we need to extract information for each aspect in the case of review summarization. Aspect summarization can present information without biasing to a specific topic. We focus on multi-aspects review summarization in this research.

Here we also focus on other information; objective information such as the market share, specifications and price of the target product. It is also important for the summarization.

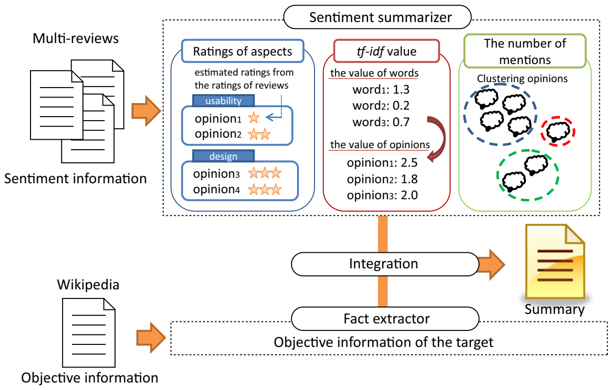


Figure 1: The outline of our method.

Integrating a sentiment summary from reviews with objective information leads to improvement in the quality of the final summary.

In this paper, we propose a method for generating a summary that contains sentiment information and objective information of a product. Figure 1 shows the outline of our method. For the sentiment summarization task, the method is based on an extraction approach. Therefore, we need to discuss which sentences are important and how to extract important sentences. In the case of treating multi-reviews, we need to handle the redundant information. Pang and Lee [10] have reported that while in traditional summarization redundant information is discarded, in sentiment summarization redundancy indicates the importance of opinions. Therefore, we treat redundancy as a feature for decision of important sentences. We assume that reviews we treat in this research have multiple aspects and a reviewer gives ratings of the aspects (i.e., 0 to 5 stars). Reviewers also write free comments about the target. We leverage three features: ratings of aspects of reviews, the $tfidf$ value, and the number of mentions with a similar topic. We apply a clustering algorithm to sentences to measure the number of mentions with a similar topic. Then, we generate a more appropriate summary by using those features. For the objective information extraction, we treat a Wikipedia entry about the current target product. We extract relative information of the product by using keywords in the sentiment summary and structural information, such as itemization, in Wikipedia.

2. Features for sentence extraction

For generating a summary, we define the following three features: (1) ratings of aspects, (2) the $tfidf$ value, and (3) the number of mentions with a similar topic. The following sections describe how we treat these features.

2.1. Ratings of aspects

If we generate a summary, the summary needs to contain the proper balance of whole opinions in the reviews. For instance, if we summarize only positive opinions, the summary cannot tell readers negative opinions of a target.

We focus on ratings of aspects given to reviews as a feature to deal with this problem. We assume that a reviewer writes comments corresponding to the ratings. For instance, if a reviewer gives 1 star for one aspect, the reviewer writes negative comments for the aspect. We assign ratings of aspects of a review to evaluative sentences included in the review. Each evaluative sentence has the rating corresponding to the aspect of the sentence. We use the pair of the evaluative sentence and the rating to consider the distribution of the ratings for summarization.

2.2. The $tfidf$ value

The $tfidf$ algorithm is used as a major algorithm for many tasks such as important sentence extraction. This algorithm features words which only appear in a target document as more important. We similarly apply the $tfidf$ algorithm to compute the importance of an evaluative sentence. First, we divide sentences in reviews to morphemes by using a morphological analyzer[1]. Then, we define the $tfidf$ value of a word i in target reviews T as below. Note that we treat only content words (except words such as suffixes and pronouns).

$$tf_i = \frac{\log_2(freq(i)+1)}{\log_2(words(T))} \tag{1}$$

$$idf_i = \log_2(\frac{Rev_{all}}{Rev_{include(i)}}) + 1 \tag{2}$$

$$tfidf_i = tf_i \times idf_i \tag{3}$$

where $freq(i)$ is the frequency of i for T, $words(T)$ is the number of words belonging to T, Rev_{all} is the number of all reviews, and $Rev_{include(i)}$ is the number of reviews including i.

Next, we compute the importance of an evaluative sentence by using the $tfidf$ value of a word. We denote an evaluative sentence by $S = \{w_1, w_2, ...\}$, where each w is a word which has $tfidf$ value in the sentence. The importance of an evaluative sentence $tfidf_S$ is as below:

$$tfidf_S = \frac{\sum_{w \in S} tfidf_w}{|S|} \tag{4}$$

where $|S|$ is the total number of w in the sentence. This is the importance of an sentence based on feature words.

2.3. The number of mentions with a similar topic

We treat multi-reviews for the target of summarization. In this case, some reviewers might write similar opinions. These similar opinions have possibilities to be redundantly extracted as a summary. On the other hand, the opinion mentioned by many reviewers is important. We need to handle redundant information regarding as a feature to determine an important sentence. Therefore, we aim to integrate similar opinions by clustering them.

In this paper, we apply the k-$means$ algorithm which is widely used as the clustering algorithm because of its simplicity. Since the k-$means$ algorithm is a non-hierarchical method, firstly we need to specify how many clusters we divide. However, It is difficult to know the optimum number of divided clusters beforehand. Seki et al. [11] have estimated the valid number of clusters by statistically evaluating the clustering result. We apply their algorithm to our task.

We divide evaluative sentences to morphemes and construct a vector space using these morphemes as features. Note that we treat content words, adjectives, and verbs. If the feature has the $tfidf$ value, a score of each feature is the $tfidf$ value computed in Section 2.2. If a feature does not have the $tfidf$ value, we simply assign the frequency within the evaluative sentence as the score. Besides, we introduce the concept of centrality of the word which has been reported by Ishii et al. [4] to characterize features. This method assumes that words such as a subject case and an objective case in a sentence indicate a topic of the sentence. We apply the concept to our task and weight central words in the evaluative sentence.

By clustering evaluative sentences based on these algorithms, we can generate some clusters including similar sentences. However, we found that clusters generated by our algorithm tended to be divided too much in a preliminary experiment. It denotes that similar opinions which should belong to the same cluster belong to other clusters. Therefore we revise the clusters by using co-occurrence of representative words of each cluster [13].

[1] We used Mecab. http://mecab.sourceforge.net/

3. Sentiment Summarizer

In this section, we describe how to generate a summary based on three features mentioned in Section 2.. We compute the importance of each cluster by integrating the $tfidf$ value of each sentence and the number of mentions. The importance of a cluster $Imp(C)$ is as below:

$$Imp(C) = Mean_{tfidf_S}(C) \times \log(|C| + 1) \qquad (5)$$

where $Mean_{tfidf_S}(C)$ is the average of $tfidf_S$ belonging to a cluster C. $|C|$ is the total number of sentences in the cluster. $Imp(C)$ is the importance which means both the importance of feature words and the number of mentions.

Besides, we treat ratings of aspects to reflect the proper balance of whole opinions of reviews. The process for the sentence extraction is as follows:

1. identify a representative sentence which has the top $tfidf$ from each cluster,
2. classify representative sentences into the rating to which the sentence belongs,
3. extract sentences with high $Imp(C)$ on the basis of the distribution of the total number of representative sentences belonging to each rating.

As the representative sentence, we select a sentence which is close to the centroid of the cluster. However, each cluster often contains sentences with different polarities; positive and negative. Therefore, we judge the major polarity of each cluster first. Then, we select the sentence which is close to the centroid and contains the polarity.

4. Integration

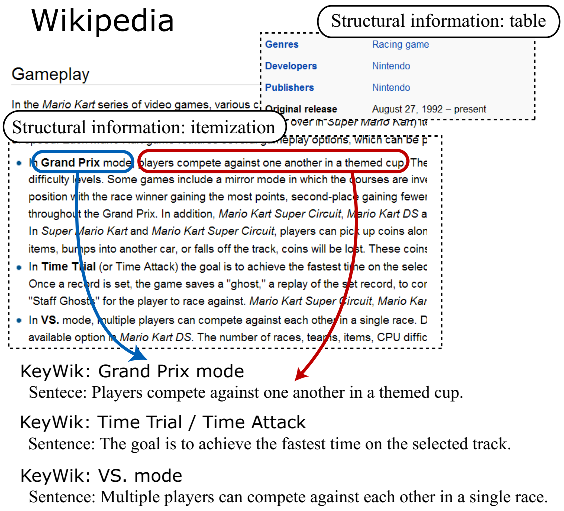
In this section, we explain an integration approach of a sentiment summary and objective information which is related to it. In this paper we regard specifications and explanations about the target product as objective information. Integrating a sentiment summary from reviews with objective information leads to improvement in the quality of the summary.

In this paper, we handle a Wikipedia entry about the target product. The integration process consists of two processes; detection and alignment of KeySum (keywords in a sentiment summary) and KeyWik (keywords in a Wikipedia entry) .

KeySum is nouns with the high $tfidf$ values in a sentiment summary. The number of KeySums in each sentence is limited by

$$\frac{\#\text{ of nouns in a sentence}}{2}$$

Our system extracts sentences including the KeySum from the Wikipedia entry. On the other hand, KeyWik is detected by using structural information, such as itemization, in Wikipedia. Figure 2 shows an example of the process. If the sentiment summary contains the KeyWik, our system aligns words in the sentiment summary and the KeyWiks.


Wikipedia
Structural information: table
Gameplay
Structural information: itemization

Genres	Racing game
Developers	Nintendo
Publishers	Nintendo
Original release	August 27, 1992 – present

In the *Mario Kart* series of video games, various c

over in *Super Mario Kart*) it

eplay options, which can be p

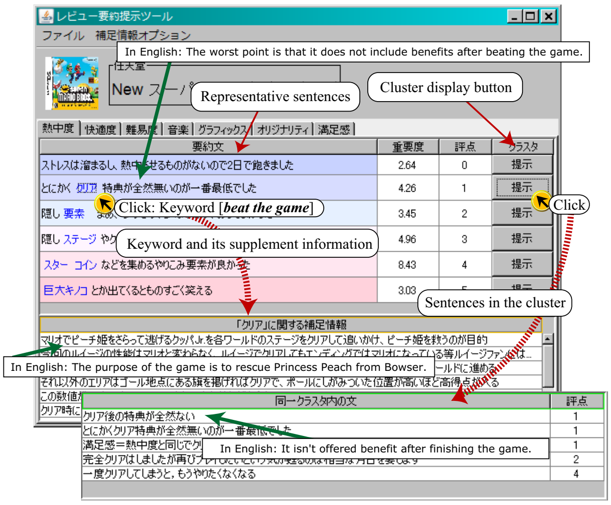
- In **Grand Prix** mode, players compete against one another in a themed cup. The difficulty levels. Some games include a mirror mode in which the courses are inve position with the race winner gaining the most points, second-place gaining fewer throughout the Grand Prix. In addition, *Mario Kart Super Circuit*, *Mario Kart DS* a In *Super Mario Kart* and *Mario Kart Super Circuit*, players can pick up coins alon items, bumps into another car, or falls off the track, coins will be lost. These coins
- In **Time Trial** (or Time Attack) the goal is to achieve the fastest time on the selec Once a record is set, the game saves a "ghost," a replay of the set record, to cor "Staff Ghosts" for the player to race against. *Mario Kart Super Circuit*, *Mario Kar*
- In **VS.** mode, multiple players can compete against each other in a single race. D available option in *Mario Kart DS*. The number of races, teams, items, CPU diffic

KeyWik: Grand Prix mode
Sentece: Players compete against one another in a themed cup.

KeyWik: Time Trial / Time Attack
Sentence: The goal is to achieve the fastest time on the selected track.

KeyWik: VS. mode
Sentence: Multiple players can compete against each other in a single race.

Figure 2: The KeyWik extraction process.



In English: The worst point is that it does not include benefits after beating the game.

要約文	重要度	評点	クラスタ
ストレスは溜まるし、熱中させるものがないので2日で飽きました	2.64	0	提示
とにかく クリア 特典が全然無いのが一番最低でした	4.26	1	提示
隠し 要素　まあ	3.45	2	提示
隠し ステージ やク	4.96	3	提示
スター　コイン などを集めるやりこみ要素が良かった	8.43	4	提示
巨大キノコ とか出てくるとものすごく笑える	3.03	5	提示

「クリア」に関する補足情報
マリオでピーチ姫をさらって逃げるクッパJr.を各ワールドのステージをクリアして追いかけ、ピーチ姫を救うのが目的
今回のルイージの性能はマリオと変わらなく、ルイージでクリアしてもエンディングではマリオになっている等ルイージファンには…
ールドに進める
それ以外のエリアはゴール地点にある旗を掲げればクリアで、ポールにしがみついた位置が高いほど高得点が入る
この数値
クリア時に

In English: The purpose of the game is to rescue Princess Peach from Bowser.

同一クラスタ内の文	評点
クリア後の特典が全然ない	1
とにかくクリア特典が全然無いのが一番最低でした	1
満足感＝熱中度と同じでクリ	1
完全クリアはしましたが再びプレイしたいという気が甦るのは相当な月日を要します	2
一度クリアしてしまうと，もうやりたくなくなる	4

In English: It isn't offered benefit after finishing the game.

Figure 3: Summary output interface.

Figure 3 shows the output of our system. The system can display (1) representative sentences in each aspect, (2) supplement information and (3) sentences belonging to each representative sentence. Each representative sentence contains the importance score ($Imp(C)$) and the rating. By clicking a keyword in the summary, our system displays sentences extracted from Wikipedia as supplement information. The supplement sentences are displayed in descending order of the importance. It is based on the concept of centrality of the word and is the same as the approach, which was proposed by Ishii et al. [4], in Section 2.3. If a user pushes a “cluster display” button, he/she can access to sentences which belong

Table 1: A summary by the proposed method for <Addiction (a)>.

Representative sentence	$Imp(C)$	$ActRat$	Ave
I'm bored with in a few days because it is stressful.	2.46	0	0.00
The worst point is that it does not include benefits after beating the game	4.26	1	1.80
The number of hidden items is good.	3.45	2	2.00
Hidden stages after completing the game are scarce.	4.96	3	3.5
I enjoyed collecting Star coins.	8.43	4	3.36
Big mushroom item is very funny idea.	3.03	5	5.00

to the cluster.

5. Experiment

In this section, we evaluated our summarization. First, we describe the data set for the experiment. Then, we compare our summary with a manual summary as quantitative evaluation. Finally, we evaluate our system with objective information qualitatively.

5.1. Data set

We used game review documents which Shimada et al. [12] used for evaluative documents classification. The review documents were extracted manually from the Web site[2]. 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)>. The review documents include reviews for 49 games. The numbers of the all reviews are 4,174 reviews. We chosen three of the Nintendo DS software[3] as the target data. They consisted of 170, 130, 24 reviews, respectively. We randomly selected approximately 450 sentences from each software. Then, three annotators (A_1, A_2, A_3) annotated the sentences. For quantitative evaluation, annotators manually generated summaries. Each annotator extracted 50 sentences as a summary from them for each software.

5.2. Evaluation of a summary

First, we discuss output summaries from our method. Table 1 shows an example of a summary for the aspect <Addiction (a)>. It contains some representative sentences and their $Imp(C)$. $ActRat$ is the actual rating of the review containing the representative sentence, "Ave" is the average of the ratings in the cluster containing the representative sentence.

The value of the Ave was close to that of $ActRat$. This result shows the effectiveness of the sentiment summarizer with polarity-adjusted sentence extraction (See Section 3.). However, one of the problems in the result of Table 1 was that there existed the sentence to which its rating did not correspond. For example, although "The number of hidden items is good." was usually considered as a positive opinion, the rating of the review containing the sentence was 2. The reason was that most of the sentences in the review containing the representative sentence had negative opinions for this aspect. This is due to the fact that opinions of reviewers sometimes might be inconsistent with the ratings. We need to handle the non-consistence of ratings if we treat the rating information.

Next, we compared our summaries with manual summaries. We used ROUGE-N [5] for evaluation of summaries. It indicates an n-gram recall between reference summaries and a candidate summary. ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in S_H} \sum_{g_N \in S} C_m(g_N)}{\sum_{S \in S_H} \sum_{g_N \in S} C(g_N)} \qquad (6)$$

where S_H is a set of reference summaries, g_N is the N length n-gram, $C(g_N)$ is the frequency of the g_N in the reference summary, and $C_m(g_N)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

Table 2 shows ROUGE-1 between our summaries and manual summaries. The baseline method was based on only the $tfidf$ value, that is a method without clustering. Those scores are the average scores in three annotators. Since the distribution of each aspect was different, we computed the weighted average Ave_{wgt}. We weighted the number of sentences in reference summary for each aspect. AnnoR denotes the weighted average ROUGE score between annotators. We regarded A_1 as the reference summary and the scores are the average of A_2 and A_3.

Even the scores between annotators were not high. This result shows that it is difficult to generate a same sentiment summary. For MC and FS, our method outperformed the baseline in most aspects and Ave_{wgt}. On the other hand, the ROUGE score of our method was slightly low as compared with that of the baseline for SM.

However, the evaluation with ROUGE has a problem. The ROUGE is based on correspondence between the output and

[2] http://ndsmk2.net/

[3] In this paper, there were "SM (New Super Mario Brothers)", "MC (Mario Cart)", "FS (Family Stadium)".

Table 2: ROUGE-N between our summaries and manual summaries.

Name	Method	a	c	d	m	g	o	s	Ave_{wgt}	AnnoR
SM	Baseline	**0.301**	**0.506**	0.357	**0.476**	0.095	**0.373**	0.335	**0.349**	0.460
	Proposed	0.275	0.453	**0.430**	0.303	**0.205**	0.341	**0.350**	0.337	
MC	Baseline	0.304	0.310	**0.468**	0.289	**0.363**	0.388	0.408	0.361	0.371
	Proposed	**0.492**	**0.439**	0.393	**0.341**	0.347	**0.456**	**0.498**	**0.424**	
FS	Baseline	0.384	0.416	0.211	0.644	**0.338**	0.371	**0.451**	0.402	0.522
	Proposed	**0.430**	**0.454**	**0.404**	0.644	0.101	**0.411**	0.441	**0.412**	

Table 3: Semantic agreement between our summaries and manual summaries (%).

Name	Method	a	c	d	m	g	o	s	Ave_{wgt}
SM	Baseline	14.3	**37.5**	21.4	25.0	33.3	27.3	20.0	25.5
	Proposed	**28.6**	31.3	**50.0**	25.0	**66.7**	**36.4**	**25.0**	**37.6**
MC	Baseline	25.0	33.3	**40.0**	0.0	20.0	31.8	22.2	24.6
	Proposed	**37.5**	**44.4**	0.0	**50.0**	20.0	**45.5**	**33.3**	**33.0**
FS	Baseline	25.0	13.6	14.3	33.3	0.0	**40.0**	**37.5**	23.4
	Proposed	**50.0**	**50.0**	**42.9**	33.3	**16.7**	30.0	29.2	**36.0**

manual summaries. Therefore, the value becomes low in the case that surface expressions are different even if the sense of sentences is similar. In this experiment, we evaluated the methods with semantic agreement between our outputs and manual summaries. First we displayed two sentences, namely our output and a manual output, to a test subject. Then, the test subject judged whether the outputs were similar in terms of content. The result shows Table 3. The number of test subjects was 2 persons. As compared with the result of Table 2, i.e., ROUGE, the proposed method generated higher scores. This result shows the effectiveness of the proposed method as compared with the baseline.

5.3. Evaluation of our system

We evaluated the effectiveness of object information and our system. This experiment was qualitative evaluation and the number of test subjects was four persons. The test subject scored 1 (bad) - 5 (good) points for each questionnaire entry. The result shows Table 4. The average score was 4.3.

In particular, the score of "effectiveness for content understanding" as a summary was high. This result shows the effectiveness of integration between object information and a sentiment summary. By using our system, a user can easily understand the product in terms of sentiment and objective information.

On the other hand, there were some negative comments from the test subjects. The 1st comment was a problem of the clustering process. The clustering process did not deal with the polarity of each sentence. Sentences with a different polarity were occasionally contained in a cluster; e.g., the polarity of a representative sentence in a cluster was positive and the cluster contained some negative sentences. To improve the clustering process is one future work. Another negative comment is concerning the keyword extraction from sentiment reviews for the integration process. There was lack of keywords for the supplement. To generate more appropriate integrated summaries, we need to consider the extraction process of target words for the supplement.

Table 4: Evaluation of our system

Effectiveness of objective information	
Correspondence between supplement information and content	4.0
Adequateness of selection of KeySum and KeyWik	4.2
Effectiveness for content understanding	4.6
Necessity of objective information	4.4
Effectiveness of our system	
As informative summary	4.2
Easy-to-understand	4.6
Average	4.3

6. Related work

Meng and Wang [7] have extracted aspects from the specification of the target product and summarized reviews with hierarchic structures. As a result, they could extract appropriate aspects for the products. However, the generated summary did not include detailed opinions about the product. In contrast, our method can treat detailed information by extracting important sentences with feature words.

Blair-Goldensohn et al. [1] have computed a polarity value of sentences based on the maximum entropy method with WordNet and rating information. They extracted evaluative sentences with a high polarity value preferentially and generated a summary. As the advantage of their method, it could estimate the polarity of sentences with high accuracy. However, it is not always true that sentences with high polarity values are appropriate for a summary. They also did not treat redundancy of the summary.

Takamura and Okumura [14] have proposed a document summarization method based on the budgeted median problem. Nishikawa et al. [8] have proposed a opinion summarization method handling content and coherence simultaneously. These methods were effective but did not deal with objective information.

Lu and Zhai [6] have introduced the concept of aspects to a PLSA model. They integrated expert reviews and ordinary opinions scattering in the Web. Opinions which should be integrated are identified by measuring the number of mentions in the Web. Although their method are very effective, their purpose is to add more information, such as similar and supplementary opinions, to the base review.

In this paper, we did not discuss the identification of the aspect of sentences. However, it is important to identify the aspect information for the sentiment summarization task. Hadano et al. [3] have identified an aspect of an evaluative sentence with machine-learning approaches. We need to introduce such a method to our task in the future.

7. Conclusion

In this paper, we focused on the multi-aspects review summarization. We handled three features; ratings of aspects, the $tfidf$ value, and the number of mentions with a similar topic. We used a clustering method to integrate similar opinions. The experimental result showed that we could integrate similar opinions and it led to the redundancy elimination of a summary.

In addition, we handled objective information for the summarization task. Supplying objective information led to improve the content understanding of a summary. The result in the qualitative evaluation showed the effectiveness of our system that integrated a sentiment summary with objective information.

Future work includes (1) more appropriate decision of representative sentences, (2) handling not only Wikipedia but also other information sources and (3) computing a confidence as objective information.

Acknowledgment

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

References

[1] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *Proceedings of WWW 2008: NLPIX Workshop*, 2008.

[2] Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[3] Masashi Hadano, Kazutaka Shimada, and Tsutomu Endo. Aspect identification of sentiment sentences using a clustering algorithm. In *Proceedings of PACLING 2011*, 2011.

[4] Hiroshi Ishii, Rihua Lin, and Teiji Furugori. An automatic text summarization system based on the centrality of word roles in sentences (in japanese). *Information Processing Society of Japan SIG Notes*, 2001(20):83–90, 2001.

[5] Chin-Yew Lin. Looking for a few good metrics: Automatic summarization evaluation – how many samples are enough? In *Proceedings of NTCIR Workshop 4*, 2004.

[6] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic. In *Proceedings of WWW 2008*, pages 121–130, 2008.

[7] Xinfan Meng and Houfeng Wang. Mining user reviews: from specification to summarization. In *Proceedings of ACL-IJCNLP 2009*, pages 177–180, 2009.

[8] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kiku. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistic (COLING '10)*, 2010.

[9] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 271–278, 2004.

[10] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2 (1-2):1–135, 2008.

[11] Tsunehito Seki, Kazutaka Shimada, and Tsutomu Endo. Acquisition of synonyms using relations in product. *Systems and Computers in Japan*, 38(12):25–36, 2007.

[12] Kazutaka Shimada and Tsutomu Endo. Seeing several stars: a rating inference task for a document containing several evaluation criteria. In *Proceedings of PAKDD 2008*, 2008.

[13] Ryosuke Tadano, Kazutaka Shimada, and Tsutomu Endo. 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)*, pages 685–692, 2010.

[14] Hiroya Takamura and Manabu Okumura. Text summarization model based on the budgeted median problem. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1589–1592, 2009.