Segmentation of Review Texts by using Thesaurus and Corpus-based Word Similarity

Yoshimi Suzuki and Fumiyo Fukumoto
Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, Japan

Keywords: Text Segmentation, Reviews, Thesaurus, Wikipedia.

Abstract: Recently, we can refer to user reviews in the shopping or hotel reservation sites. However, with the exponential growth of information of the Internet, it is becoming increasingly difficult for a user to read and understand all the materials from a large-scale reviews that is potentially of interest. In this paper, we propose a method for review texts segmentation by guest's criteria, such as service, location and facilities. Our system firstly extracts words which represent criteria from hotel review texts. We focused on topic markers such as "ha" in Japanese to extract guest's criteria. The extracted words are classified into classes with similar words. The classification is proceeded by using Japanese WordNet. Then, for each hotel, each text with all of the guest reviews is segmented into word sequence by using criteria classes. Review text segmentation is difficult because of short text. We thus used Japanese WordNet, extracted similar word pairs, and indexes of Wikipedia. We performed text segmentation of hotel review. The results showed the effectiveness of our method and indicated that it can be used for review summarization by guest's criteria.

1 INTRODUCTION

Recently, we can refer to user reviews in the shopping or hotel reservation sites. Since a user's criterion is estimating the user review compared with the information which a contractor offers, there is a possibility that many information which is not included in a contractor's explanation is included. These customer/guest reviews are often included various information about products/hotels which are different from commercial information provided by sellers/hotel owners, as customer/guest have pointed out with their own criteria, *e.g.*, service may be very important to one guest such as business traveler whereas another guest is more interested in good value for selecting a hotel for his/her vacation. However, there are at least seven problems as follows:

- There is a large amount of reviews for each product/hotel.
- Each review text is short.
- There are overlapping contents.
- The wrong information may be described.
- It is not faithful to grammar.
- The terms are not unified.
- There are many miss spellings, words.

Moreover, it is difficult to find boundary for every item. Because reviews explain one item by using two or more sentences, or two or more items are explained by using only one review sentence.

In this paper, we propose a method for review texts segmentation by using guest's criteria, such as service, location and facilities.

2 RELATED WORK

Text segmentation is one of the challenging tasks of Natural Language Processing. It has been widely studied and many techniques (Kozima, 1993) (Hearst, 1997) (Allan et al., 1998) have been proposed. Hearst presented a method of TextTiling (Hearst, 1997), which is based on lexical cohesion concerning to the repetition of the same words in a document. Utiyama and Isahara proposed a statistical method for domain-independent text segmentation (Utiyama and Isahara, 2001). Hirao et al. proposed a method based on lexical cohesion and word importance (Hirao et al., 2000). They employed two different methods for text segmentation. One is based on lexical cohesion considering co-occurrences of words, and another is base on the changes of the importance of the each sentence in a document.

Suzuki Y. and Fukumoto F..
Segmentation of Review Texts by using Thesaurus and Corpus-based Word Similarity.
DOI: 10.5220/0004140903810384
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 381-384
ISBN: 978-989-8565-30-3

3 SYSTEM OVERVIEW

Figure 1 illustrates overview of our system. The system consists of two modules, namely “Extraction of criteria”, “Text segmentation”.

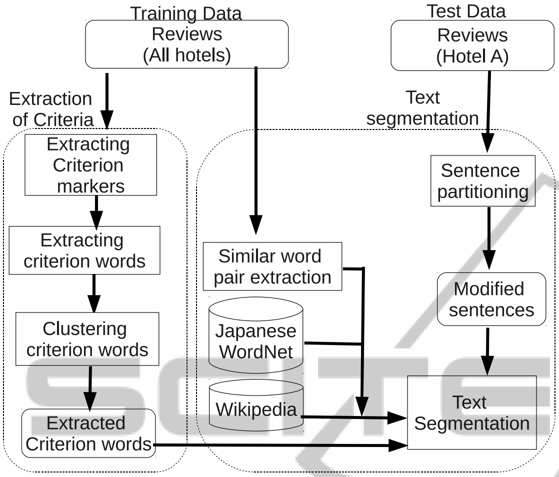

Figure 1: System overview.

4 EXTRACTION OF CRITERIA

Firstly, we define criteria as words which reviewers notice in the reviews. In the text in Japanese, criteria are represented followed by “ha”. For extracting criteria in reviews, firstly we extract postpositional particle “ha” from whole review texts. Next we extract words followed by “ha”, and finally, we collected words which are index words of Japanese WordNet from the extracted words as criteria words. Table 1 shows extracted criteria words with high frequency.

Table 1: Candidate words of criteria (top 5).

No	words	frequency
1	room	56,888
2	breakfast	25,068
3	meal	17,107
4	support	16,677
5	location	14,866

5 SIMILAR WORD PAIR EXTRACTION

Review texts are written by many different people. People may express the same thing by using different expression. For example, “*heya*”, “*oheya*” and “*ruumu*” are the same sense, *i.e.*, room. Moreover, two words such as “*kyakushitsu*”:(guest room) and “*heya*”:(room) are often used in the same sense in the hotel review domain while those are different senses.

We thus collected similar words from hotel reviews by using Lin’s method (D.Lin, 1998).

Some technical terms do not frequently appear in reviews even if we use large corpora. Therefore, we applied smoothing technique to the hierarchical structure of semantic features. Firstly, we extracted similar word pairs using dependency relationships. Dependency relationship between two words is used for extracting semantically similar word pairs. Lin proposed “dependency triple” (D.Lin, 1998). A dependency triple consists of two words: w, w' and the grammatical relationship between them: r in the input sentence. $||w,r,w'||$ denotes the frequency count of the dependency triple (w,r,w'). $||w,r,*||$ denotes the total occurrences of (w,r) relationships in the corpus, where “$*$” indicates wild card.

We used 3 sets of Japanese case particles as r. Set A consists of 2 case particles: “ga” and “wo”. They correspond subject and object, respectively. Set B consists of 6 case particles. Set C consists of 17 case particles. We selected word pairs which are extracted by using two or three sets.

In order to extract the corresponding semantic feature of the new word, we extracted dependency triples of the new word and the extracted words. Using some extracted words, many types of dependency triples are extracted. For extracting the similar words from the core thesaurus, $I(w,r,w')$ is calculated by using Formula (1).

$$\begin{aligned} I(w,r,w') & \\ &= -\log(P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)) \\ &\quad -(-\log P_{MLE}(A,B,C)) \\ &= \log\frac{||w,r,w'||\times||*,r,*||}{||w,r,*||\times||*,r,w'||} \qquad (1) \end{aligned}$$

where P_{MLE} refers to the maximum likelihood estimation of a probability distribution.

Let $T(w)$ be the set of pairs (r,w') such that $\log\frac{||w,r,w'||\times||*,r,*||}{||w,r,*||\times||*,r,w'||}$ is positive. The similarity $Sim(w_1,w_2)$ between two words: w_1 and w_2 is defined by Formula (2).

$$\begin{aligned} & Sim(w_1,w_2) \\ &= \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1,r,w)+I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w) + \sum_{(r,w)\in T(w_2)} I(w_2,r,w)} \qquad (2) \end{aligned}$$

Table 2 shows the extracted pairs as similar word pairs.

In Table 2, there are some notational variants. In general, the pair of “morning newspaper” and

Table 2: Results of extracting similar pairs using particle set A, B, C.

No.	word1	word2
1	favorable (*koukan*)	very favorable (*taihen koukan*)
2	route (*michizyun*)	route (*ikikata*)
3	stomach (*onaka*)	stomach (*onaka* hiragaga)
4	dust (*hokori*)	dust (*hokori* hiragana)
5	net (*netto*)	Internet (*intaanetto*)

"newspaper" and the pair of "breakfast voucher" and "ticket" are not same meaning, however the two pairs are mostly same meaning in hotel review texts.

6 SENTENCE PARTITIONING AND TEXT SEGMENTATION

Compound sentences frequently appear in review texts. Moreover, two or more criteria may be included within a compound sentence. For example, "The buffet-style breakfast is delicious, the room is also large and the scent of the shampoo and rinse in the bathroom are quite good": "(*chooshoku no baikingu mo oishiidesushi, heyamo hiroishi, ichiban kiniitteiruno ga heya ni oitearu shampuu to rinsu no kaori ga totemo iito omoimasu*)".

It is necessary to divide one sentence into some criteria. Fukushima proposed a method of sentence division for text summarization for TV news (Fukushima et al., 1999). They used rule based method for sentence partitioning. In this paper, each compound sentence was divided into some criteria by using compound sentence markers and "cabocha" ((Kudo and Matsumoto, 2002)) which is a Japanese dependency structure Analyzer.

Using results of sentence partitioning, we divided text by criteria using lexical information of Japanese WordNet, similarity of words and indexes of Wikipedia. Figure 2 shows how to divide text.

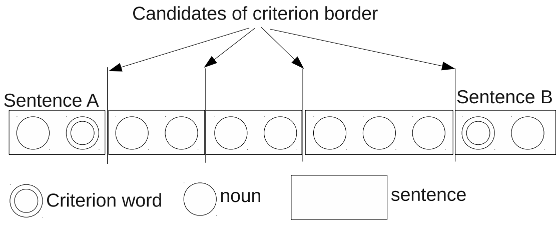

Figure 2: Text segmentation method.

We decide border of criteria between 2 different criterion words. The following algorithm is how to decide border of criteria:

```
for(i=0; i<N; i++)
{
  if(similarityA[i]<similarityB[i]) then
    break;
}
border=i;
```

where N is number of sentences between two sentences which have criterion words. similarityA[i] and similarityB[i] are similarity between *i*th sentence and "sentence A" and "sentence B", respectively. Each similarityA[i] is calculated by following algorithm:

```
similarityA[i]=0.0
for(j=0; j<M; j++)
{
  similarityA[i]+=
  cooccurrence(criterionA,wordj)
  +cooccurrence(criterionA,simpair(wordj))/2
  +cooccurrence(criterionA,hypernym(wordj))/2
  +cooccurrence(criterionA,wikipedia(wordj))/2;
}
```

where M is number of words in sentence[i]. "criterionA" is the criterion in "sentence A". "wordj" is *j*th word in sentence[i]. "simpair(wordj)" is the word of similar word pair of wordj. "hypernym(wordj)" is the hypernym of wordj. "wikipedia(wordj)" is the most right noun of first sentence of wordj. "occurrence(x,y)" is the rate which x and y are co-occurring in same sentence of the hotel review.

7 EXPERIMENTS AND DISCUSSION

For the experiment, we used hotel review of Rakuten Travel[1]. Table 3 shows Review data of the Rakuten Travel.

Table 3: Reviews of Rakuten Travel.

amount of data	250MB
# of review text	350,000
# of hotel	15437
# of words for each review	375
# of reviews for each hotel	23

We used Japanese WordNet Version 1.1 (Bond et al., 2009) as Japanese Thesaurus dictionary. Also we used index words of Wikipedia[2] for dealing with

[1] url= http://travel.rakuten.co.jp/ We used Rakuten travel review data provided by Rakuten Institute of Technology

[2] url=http://ja.wikipedia.org/ We used Wikipedia data of 2012-03-26.

words such as named entities which are not index words of Japanese WordNet. We employed Lin's method (D.Lin, 1998) for extracting similar word pairs in hotel review texts.

We had experiments for dividing reviews into every criterion. We used review texts of 5 hotels. The average number of review texts per hotel was 51.2. The number of criteria consists of 256. Table 4 shows the results of text segmentation.

Table 4: Results of text segmentation.

	Our method	TextTiling
Precision	863/1024=0.842	742/1024=0.725
Recall	863/902=0.925	742/1119=0.663
F-measure	0.882	0.692

We compared our method with TextTiling (Hearst, 1997) which is a well known text segmentation technique. TextTiling is a technique for subdividing texts into multi-paragraph units that represent passages, or subtopics. The discourse cues for identifying major subtopic shifts are patterns of lexical co-occurrence and distribution. TextTiling shows high performance/accuracy for documents which have rather long topics such as magazine articles. However, TextTiling could not obtain good results as review text for each criterion of is very short. As can be seen clearly from Table 4, the results obtained by our method were much better than those of TextTiling. This demonstrates that lexical information such as similarity between words, hypernyms of words and named entities were effective for text segmentation.

8 CONCLUSIONS

In this paper, we proposed a method for review texts segmentation by guest's criteria, such as service, location and facilities. The results showed the effectiveness of our method as the results attained at 0.84 precision, 0.92 recall, and 0.88 F-measure, and it was outperformed the results obtained by TextTiling which is used for text segmentation. Future work will include: (i) applying the method to a large number of guests reviews for quantitative evaluation, (ii) applying the method to other data such as grocery stores: LeShop[3], TaFeng[4] and movie data: MovieLens[5] to evaluate the robustness of the method.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their comments on the earlier version of this paper. This work was partially supported by The Telecommunications Advancement Foundation.

REFERENCES

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study final report. In *the DARPA Broadcast News Transcription and Understanding Workshop*.

Bond, F., Isahara, H., Uchimoto, K., Kuribayashi, T., and Kanzaki, K. (2009). Enhancing the japanese wordnet. In *The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP*.

D.Lin (1998). Automatic retrieval and clustering of similar words. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference*, pages 768–774.

Fukushima, T., Ehara, T., and Shirai, K. (1999). Partitioning long sentences for text summarization. *Journal of Natural Language Processing (in Japanese)*, 6(6):131–147.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. In *Association for Computational Linguistics*, pages 111–112.

Hirao, T., Kitauchi, A., and Kitani, T. (2000). Text segmentation based on lexical cohesion and word importance. *Information Processing Society of Japan*, 41(SIG3(TOD6)):24–36.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31th Annual Meeting*, pages 286–288.

Kudo, T. and Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. In *CoNLL 2002:Proceedings of the 6th Conference on Natural Language Learning 2002*, pages 63–69.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 499–506.

[3] www.beshop.ch

[4] aiia.iis.sinica.edu.tw/index.php?option=com_docman&task=cat_view&gid=34&Itemid=41

[5] http://www.grouplens.org/node/73