Searching a Best Product Based on Mining Comparison Sentences

Yeong Hyeon Gu[1] and Seong Joon Yoo[1]
[1]Department of Computer Engineering, Sejong University
Gunja, Gwangjin, Seoul, Korea
Email: sjyoo@sejong.ac.kr

***Abstract*—Research on extracting technology of structure information from comparison sentences are to extract comparison target, features, and relations from given sentences and it is a very important technology in comparing search. There have been studies in English speaking countries, but this is the first study on Korean documents. In this study, the comparison targets and relations were focused among the structure information of comparison sentences to find the superiority and inferiority determination rules. Then, the comparison keywords for superiority and inferiority as well as the predicate were used to extract superior targets among the comparison targets. Terms ''boda (than)', 'bihae (compared with)', 'hweolssin (far)', and 'deo 'more' were used for experiment and 93.9% of accuracy was resulted.**

I. INTRODUCTION

In the review documents that contain user's opinion, people set a comparative model to compare. For example, a sentence "A monitor is better than B monitor" shows that a user prefers the A monitor than B monitor therefore positive opinions can be resulted about the A monitor. By comparing various features of a specific product and categorizing positive and negative opinions for each feature, comprehensive opinions can be provided for a product and it can respond to the comparative search for the product. Studies on such 'Comparative Sentence Mining' can be applied to various fields. For companies, they can find what to improve by analyzing the users' opinions on products and services. For individuals, being able to see huge amount of systematically organized and analyzed reviews on webs rather than a small amount of reviews gathered by a portal site will help them to make decision of purchasing and this can be utilized as the opinion mining factor technologies [1]-[6].

There are many studies in comparative sentence classification methods in English speaking countries, and recently there are studies applied to Korean. However, there have no research on comparison targets and superior targets which are very useful information in comparative search in Korean so more studies on broad field and more researchers are required.

The methods to find targets that are in superior position than the comparing target from different levels of comparison sentences by using rules are described in this study.

Chapter 2 describes the related works and chapter 3 describes the extracting method of superior targets from comparative sentences by using the comparative words and sentence pattern rules. Finally chapter 4 describes the results acquired from experiments.

Corresponding author: Seong Joon Yoo (e-mail: sjyoo@sejong.ac.kr)

II. RELATED WORK

The comparative document classifying skill, which belongs to the realm of a study on opinion mining, can be said to utilize opinion mining – which uses some rule or mechanical learning bases method - in a broad sense. Opinion mining [1]-[7] finds useful information out of a massive group of information such as reviewed data. Its representative application is the summarization of product reviews. Studies classifying sentences including comparative opinions using some keywords in English speaking world include [8], [9]. In this study, 83 keywords were identified and used in those sentences. Most of them are comparative adjectives, superlative adjectives, comparative adverbs and superlative adverbs. 'more', 'less', 'most', and 'least' are not this case along with a verb such as 'beat', 'exceed', or 'ahead', which do not belong to the said part of speech but used as keywords expressing comparative opinions, were also categorized into comparative keywords. This classification of sentences of comparative opinions using keywords indicated high Recalls (98%) but very low Precisions (32%). To solve this problem, the Naive Bayes classifier was used one more time; CSR(Class Sequential Rule)[10] was utilized by virtue of the attributes of the Naïve Bayes but this was also found to be insufficient so 13 manually-established Rules were added.

Suggested in [8], [9] are not only the extraction of comparative opinion sentences but also a method for classifying the sentences into three types including non-equal gradable, equative, and superlative ones. This study has adopted various kinds of classifiers for the classification; of them, SVM[11], [12] showed the best result. With respect to Features, patterns were used ; if the sentence includes a pattern, the corresponding value was set to 1, otherwise, 0. In [13], components of a comparative sentence were analyzed to extract the subject of comparison and its properties as well as elements with a comparative edge. For this purpose, the comparative sentence is divided into 4 types and rules appropriate for each type are applied.

[14] is about application of [8], [9] tailored to Chinese language. Comparative sentence were classified using the

mechanical learning after the analysis of comparative Chinese sentences and sequential patterns in those sentences.

[15] is about sentence classification of components of comparative Chinese sentences such as a subject, properties of comparison or emotion. Semantic Role Labeling(SRL) was used to identify 6 types of components in comparative sentences; the learning and classification were implemented through Conditional Random Fields(CRFs).

[16] is about classification of superlative sentences, one of comparative sentences using Combinatory Categorical Grammar and Discourse Representation Theory. In [17] product reviews were collected and analyzed for comparison of their various properties. Property and emotional words were collected using opinion keywords and the emotion was measured in a numeric representation for each if its properties. In the numeric representation of emotion, the adverbs modifying the emotional word were divided into 4 categories and the weights were differentiated. Using those values made it easier to identify the differences between different products or their properties.

[18] is summarization of characteristics of vocabularies used in modern comparative Korean phrases, of comparison and metaphor and of the comparative, and the system of comparative phrases, properties of equality and graded comparisons. In [19], 177 comparative vocabularies that can be used as a comparative phrase in comparative Korean sentences were identified. This classification of sentences using these words indicated high Recalls but very low Precisions, and the classification was implemented again using the mechanical learning. Unlike [8], [9], the Sequential Rule was not used as a Feature of mechanical learning, and instead the outcome of analyzing the morpheme was directly used. As a mechanical learning algorithm, Naïve Bayes and Maximum Entropy Model were used; the test result revealed a higher comparative performance for Maximum Entropy Model. [20] focuses on 'boda (than)' preferred the most in comparative Korean sentences of all comparative vocabularies to identify some rules for distinguishing comparative sentences. The Korean word 'boda' is a proposition which has the same role as the English word, 'than'; if used as a adverb, 'more'. In total, 11 rules were found in the observation of product reviewing documents collected from the web using the word 'boda'. There are many studies in English speaking documents on comparative sentences classification methods and there has been some studies applied to Korean recently. However, studies on comparative targets and superior targets which can be very useful information for comparative search have not applied to Korean.

III. EXTRACTION OF SUPERIOR TARGETS FROM COMPARISON SENTENCES

Comparative sentences have certain structures. Those are comparing subject, comparing target, comparing features, and comparing superiority. For example, a sentence "Restaurant A is more delicious than restaurant B" was examined. Restaurant A is a comparing subject and restaurant B is a comparing target. Although it was not specifically described here, the flavor of food is a comparing feature and A restaurant is actually superior target than restaurant B.

Korean comparison sentences are classified into equality and inequality comparison upon the first judgment for the similarity and difference of two targets. Equality comparison is divided into simple equality and similarity/equality comparison upon its degree. The inequality comparison is divided into simple inequality and degree inequality comparison upon its degree. Degree inequality is divided into superiority/inferiority comparison and superlative upon the degree of the comparison target and superiority/inferiority comparison is divided into superiority and inferiority comparison and superlative is divided into absolute superlative and relative superlative. The comparison words that are used specifically for comparative expression in Korean are as follows.

The words used for simple equality comparison are predicates as '같다(same)', '동일하다(equal)', '동등(대등)하다(identical)', '마찬가지다(not different)', '일치되다(not discordant)', '한가지다(like kind)' and similar equality predicates are '비슷하다(similar)', '유사하다(alike)', '흡사하다(almost same)', '닮다(resembling)', '혹사하다(resemble closely)', '버금가다(in the same manner)' and propositions are '같이(as same as)', '처럼(like)', '만(only)', '만큼(as~as)'. Also the simple graded comparison words are '다르다(different)', '상이하다(unlike)', '차이가 있다(distinct)', '대조적이다(contrasting)', '대립되다(conflict)', '판이하다(dissimilar)' and propositions for superiority/inferiority words are '보다(than)', '에서(in)', '가운데(among)', '비하여(비해서)(compared and compare to)' and adverbs are '잘(well)', '더(more)', '덜(less)', '못(poor)', '훨씬(far)'. The superior words are '제일(most)', '으뜸(best)', '최고(highest)', '최상(the top)', '최저(lowest)', '최하(worst)', '가장(greatest)'.

In this study, the words used for graded comparison among the above words are used and the results of studying how to determine the superior targets on the comparison are summarized.

In order to collect the comparison sentences, the web crawler is used to collect comments from restaurant information provision websites.

Comments from the websites providing restaurant information as well as restaurant lists were collected. It is because comparison targets and subjects can be easily extracted later with the restaurant lists. Since Korean has unregulated order of working and the subject is frequently omitted, there are many sentences without the subject. In this case, comparison subject information couldn't be found in a sentence, the title of a posting or restaurant information included in the document, or overall document need to be reviewed. This is complicated and makes it difficult to analyze phrases, restaurant lists were used to extract the candidates of the subject and comparison targets.

Then, restaurant lists collected from websites providing restaurant information were used to classify the users' comments on restaurants as well as the comments with the

name of the restaurant mentioned. It was because the comments that included other restaurants' name rather than the restaurants that users wrote a comment about can be compared to other restaurants.

Also sentence structure rules and statistic methods were used to extract comparative sentences from comments. In this model sentences are retrieved from the document and then those sentences containing any comparative words are extracted. The extracted sentences are classified into comparative sentences or non-comparative ones by the Rule based comparative sentence analyzer, with comparative ones added to the database. Non-comparative ones go through another classification procedure provided by the analyzer, with comparative ones added to the database.

In this thesis, only the graded comparison sentences are targeted, so only the words which are used in the graded comparison sentences such as 'than', 'more', and 'far' are targeted.

In the result of analyzing the graded comparison sentences to find out the superior targets of comparison sentences, a few types are found. First of all, Type 1 is when there is only one comparison target in a comparison sentence. Oppositely, type 2 is when there are many comparison targets in a comparison sentence.

In order to section the comparison targets, propositions, commas (','), and conjunctions are used in Type 2. For example, propositions such as 'gwa/wa', 'na/yina', and 'wa hamgye' can be used to list out comparison targets; conjunctions such as 'or', and 'and' can be used; also comma (',') can be used. Therefore, propositions, conjunctions, and commas are used for grouping the comparison targets. Then, those are compared with the comparison targets to determine the superior targets.

In order to determine the superior targets from superiority/inferiority comparison sentences, the predicate information that is located closest next to the superiority/inferiority comparison word. At this time, Type 1 and 2 can be categorized into two cases as below. One is using propositions such as 'then(boda)' and 'compared to(bihae)' as superiority/inferiority comparison words and another one is using adverbs such as 'more(deo)' and 'far(hwelsin)' as the superiority/inferiority comparison words. The reason of categorizing into two cases is because there is a difference in categorizing superior targets. In order to find out superior targets by using predicates, a predicate (adjective or verb) that is most closely located in the right side from the comparison word is found. At this time, if the predicate that is most closely located from the comparison word in a sentence used superiority/inferiority comparison words is positive word, the comparison subject becomes the superior target and if it is a negative word, the comparison target becomes the superior target. Oppositely, if the comparison word is used for adverb and if the predicate is positive, the comparison target becomes the superior target and if it is negative, the comparison subject becomes the superior target. This can be summarized as Table 1.

TABLE I
RULES TO DETERMINE SUPERIOR TARGET FROM DECLARATIVE SENTENCES

Category		Details
Sentences using superiority /inferiority comparison words as proposition	Comparative words	'Than', 'Compared to' etc
	POS	Comparison target + than/compared to/compare to +adjective
	Rule	If the predicate is positive, → The comparison subject becomes the superior target. If the predicate is negative, → The comparison target becomes the superior target.
	Example	"Personally, this house is considered better than Umaido." "Staff are more kind than Myeongdong Gyoja where I got occasionally. "
Sentences using superiority /inferiority comparison words as adverb	Comparative words	'More', 'Far', etc
	POS	Comparison target + Proposition + Adverb + Adjective
	Rule	If the predicate is positive, → The comparison target becomes the superior target. If the predicate is negative, → The comparison subject becomes the superior target.
	Example	"Bochoen that is in across taste more good." "I heard that Giggu in East Yichon-dong is famous, but I liked Moris far better. "

In order to analyze word classes, the morpheme analyzer was used[21]. When the morpheme analyzer was used to find a predicate, it analyzed adjectives and verbs well, however, predicates such as 'Chinjelhada(kind)' and 'Chucheonhada(recommend)' were analyzed as 'noun + predicate suffix + ending'. Instead of analyzing Chinjelhada(kind)' and 'Chucheonhada(recommend)' as adjective or verb as it was in English, it analyzed as noun for like 'Chinjel' and 'Chucheon'. In order to solve this problem, when there is predicate suffix and ending is added behind a noun, it is considered as a predicate.

Since the positivity and negativity of a predicate is upon the comparison features, so the meaning actually has to be changed upon the comparison features. However, not only the comparison subject but also the comparison features are omitted in Korean, and since not many domains are dealt in this study, the comparison features are not considered in this work.

Positivity and negativity of predicates are made as a list and saved in DB. Table II shows the examples of positive and negative predicates.

For general declarative sentences, superior target could be found by using comparison words and positivity and negativity of predicates, however, for negative sentences with negative words, the superior target needs to be determined oppositely from the previous rule. For example "A is not better than B", the proposition 'than(boda)' was used as a comparison word and since there is a positive predicate 'better(jota)', the comparison subject A is supposed to be superior according to the rule, however, actually comparison target B is the superior

target. Likewise, if a negative word ‘not’ is followed after a predicate, the superior target needs to be determined oppositely from the rule for general declarative sentence, as shown in Table III.

TABLE II
EXAMPLES OF POSITIVE AND NEGATIVE PREDICATE LISTS

Category	Example
Positive Predicate	V_높, V_낫, N_친절, V_낫, V_맛있, N_저렴, V_싸, V_좋, V_괜찮, N_만족, N_유명, N_고급, V_크, V_많, N_추천, etc
Negative Predicate	V_없, V_떨어지, V_못하, N_부족, V_비싸, V_느끼하, N_한산, V_안좋, V_적, V_덜하, V_좁, V_못되, etc

TABLE III
RULES TO DETERMINE SUPERIOR TARGET FROM NEGATIVE SENTENCES

Category		Details
Sentences using superiority /inferiority comparison words as proposition	Comparison words	‘Than’, ‘Compared to’ etc
	POS	Comparison target + than/compared to/compare to +adjective+않다/없다(not)
	Rule	If the predicate is positive, → The comparison target becomes the superior target. If the predicate is negative, → The comparison subject becomes the superior target.
Sentences using superiority /inferiority comparison words as adverb	Comparison words	‘More’, ‘Far’, etc
	POS	Comparison target + Proposition + Adverb + Adjective+ 않다/없다(not)
	Rule	If the predicate is positive, → The comparison subject becomes the superior target. If the predicate is negative, → The comparison target becomes the superior target.

IV. EXPERIMENT

For experiment, web crawler was used for restaurant review websites Menupan.com and Wingbus to retrieve users’ comments and basic information such as name of restaurants and phone numbers. A total of 3,984 restaurant information were retrieved from Menupan.com and Wingbus without including the duplications. In order to avoid duplication issues, phone numbers are used instead of name of restaurants to check the duplication and it is because spaces or name spellings may different by websites. The comments collected by using the web crawler was 34,978 from Wingbus and 15,903 from Menupan.com and the total was 50,881.

From the comment collected, the comments mentioned other restaurant names other than actual commenting restaurant were found. In the result, a total of 120,329 comments were found and the reasons why there were so many results found were first, because many restaurant names could be mentioned in one comment. For example, in a sentence like “I think B or C restaurant are better than A restaurant”, instead of comparing one restaurant with another restaurant as 1:1, multiple restaurants were compared, multiple results could be made. Also some comments used general noun or pronoun were used instead of the actual name of the restaurant, and almost all comments were this cases, so it brought more results. The representative examples are using “here”, “club”, or “rice”. Since using general noun or pronoun for a few restaurants can affect the whole data, these restaurants were excluded then the comments mentioned other restaurant names were found again.

In order to exclude general nouns or pronouns, the morpheme analyzer was used. The comments that used proper nouns instead of general nouns or pronouns are found through the morpheme analyzer by inputting restaurant names. In the result of searching again after this process, a total of 11,476 comments were found. However, because a few restaurants used menu names such as ‘sushi’, ‘pork-cutlet’ or ‘handmade noodle’ in their business name, totally unrelated comments were included. In order to solve this problem, 30 restaurants that used food name in their business name were excluded from the search keyword list to search. In the result, a total of 3,959 filtered results were retrieved.

299 Comparison sentences included the word ‘than (boda)” among three data and superior comparison targets were found in 275 sentences which showed 91.9% of precision. 122 Comparison sentences included the word ‘compared to (bihae)” and superior comparison targets were found in 100 sentences which showed 81.9% of precision. 30 Sentences included the comparison word ‘far(Hwelsin) and all of them found the superior comparison target. Also 117 Sentences included the comparison word ‘more(deo)’ but 4 sentences couldn’t find the superior comparison target which showed 96.5% precision.

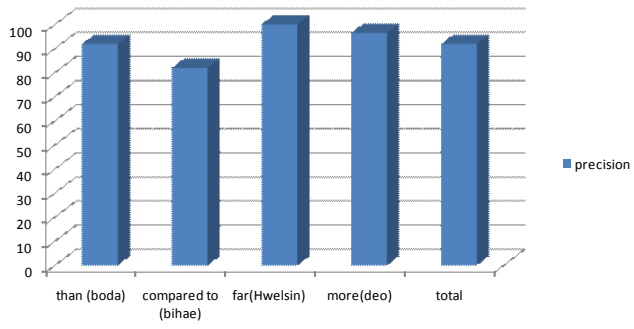

Fig. 1. Precision of Determining the Superior Targets

In the result of examining the sentences that were incorrectly analyzed or couldn’t induce superior targets, most of them were due to the incorrect spellings. In addition, the sentences used adnominal phrase or complex sentences(compound sentence) couldn’t be categorized with above defined rules.

V. CONCLUSION

In this work, the methods to extract comparison superior targets in superiority/inferiority comparison sentences by using comparison words and sentence pattern rules are studied. In

order to find the comparison subjects and targets, restaurant lists are compiled and the comments that may include comparison sentences were found by using the restaurant list, and then classified the comparison sentences through the method that integrated and applied the sentence structure rules and statistic methods. Then, the superiority/inferiority comparison keywords and predicates were used to extract the superior targets. The words 'than', 'compared to', 'far', and 'more' were used for experiment, and 91.9% of precision was resulted.

In the future, more comparison words need to be used for experiment, also various rules derived from general rules have be to found also the extraction methods for superior target information from adnominal phrases and complex sentences need to be studied. Moreover, the methods to process sentences that can't determine the superior target from will be studied.

ACKNOWLEDGMENT

This work was supported by the Korea Research Foundation(KRF) grant funded by the Korea government(MEST) (No. 2010-0015842)

REFERENCES

[1] Y. Gu and S. Yoo, "Mining Comparative Sentences from Korean Text Documents Using Sentential Structure Analysis Combined with Machine Learning Techniques"International Conference of Computer Science and its Applications(CSA 2009), Vol.1, pp.171-176 December 2009

[2] B. Pang , L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," Proc. the Conference on Empirical Methods in Natural Language Processing, pp.79-86, 2002

[3] K. Dave, S. Lawrence, and D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," In Proc. Of the 12th Intl. World Wide Web Conference(WWW '03), pp. 512-528, 2003

[4] B. Liu, "Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data," Springer

[5] D. Peter, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proc. the 40th Annual Meeting of the Association for Computational Linguistics, pp.417-424, 2002

[6] J. Hwang and Y. Ko , "A Korean Sentence and Document Sentiment Classification System Using Sentiment Features," Journal of Korean Institute of Information Scientists and Engineers (KIISE): Computing Practices and Letters, Vol. 14, No. 3, pp. 336-340, May 2008. (ISSN 1229-6848)

[7] Y. Kim, Y. Jung, and S. Myaeng, "An Opinion Analysis System Using Domain-Specific Lexical Knowledge," Proc. the 4th Asia Information Retrieval Symposium, AIRS 2008, LNCS 4993, pp.466-4.

[8] N. Jindal and B. Liu, "Mining Comparative Sentences and Relations," AAAI'06, 2006.

[9] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Document," Proc. the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 06), 2006.

[10] M. Hu and B. Liu, "Opinion Feature Extraction Using Class Sequential Rules," the 2006 AAAI Spring Symposium, pp.61-66, 2002

[11] T. Mitchell, Machine learning, McGraw-Hill, 1977.

[12] T. Joachims, Making large-scale SVM learning practical. Advances in Kernel Methods-Support Vector Learning, B.Schlkopf and C. Buges and A. Smola(ed.), 1999.

[13] B. Liu, M. Ganapathibhotla, "Mining Opinion in Comparative Sentences", Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp.241–248, Manchester, August 2008.

[14] X. Huang, X. Wan, J. Yang, J. Xiao "Learning to Identify Comparative Sentences in Chinese Text", Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence, Vol.5351, pp.187 – 198, December 2008

[15] F. Hou, G. Li, "Mining Chinese Comparative Sentences by Semantic Role Labeling", Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, Vol.5, pp.2563-2568, July 2008.

[16] J. Bos, M. Nissim, "An Empirical Approach to the Interpretation of Superlatives", EMNLP 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp 9-17, 2006.

[17] J. Sun, C. Long, X. Zhu, M. Huang, "Mining Reviews for Product Comparison and Recommendation", Research journal on Computer science and computer engineering with applications, pp.33-40, January 2009.

[18] G. Ha, A Study of Modern Comparative Syntax of Korean, PJbook, 1999.

[19] Y. Ko and Y. Ko, "Extracting Comparative Sentences from Korean Text Documents Using Comparative Lexical Patterns and Machine Learning Techniques," Proc. the ACL-IJCNLP 2009, pp.153-156, August 2009.

[20] Y. Gu and S. Yoo, "Rules for Mining Comparative Online Opinions,"International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2009), pp.1294-1299, November 2009.

[21] S. Gang, Analysis of Korean Morphemes and Information Retrieval. Hungrung Publish, 2002.