Extracting Comparative Sentences from Korean Text Documents Using Comparative Lexical Patterns and Machine Learning Techniques

Seon Yang
Department of Computer Engineering,
Dong-A University,
840 Hadan 2-dong, Saha-gu,
Busan 604-714 Korea
syang@donga.ac.kr

Youngjoong Ko
Department of Computer Engineering,
Dong-A University,
840 Hadan 2-dong, Saha-gu,
Busan 604-714 Korea
yjko@dau.ac.kr

Abstract

This paper proposes how to automatically identify Korean comparative sentences from text documents. This paper first investigates many comparative sentences referring to previous studies and then defines a set of comparative keywords from them. A sentence which contains one or more elements of the keyword set is called a comparative-sentence candidate. Finally, we use machine learning techniques to eliminate non-comparative sentences from the candidates. As a result, we achieved significant performance, an F1-score of 88.54%, in our experiments using various web documents.

1 Introduction

Comparing one entity with other entities is one of the most convincing ways of evaluation (Jindal and Liu, 2006). A comparative sentence formulates an ordering relation between two entities and that relation is very useful for many application areas. One key area is for the customers. For example, a customer can make a decision on his/her final choice about a digital camera after reading other customers' product reviews, e.g., "Digital Camera X is much cheaper than Y though it functions as good as Y!" Another one is for manufacturers. All the manufacturers have an interest in the articles saying how their products are compared with competitors' ones**.**

Comparative sentences often contain some comparative keywords. A sentence may express some comparison if it contains any comparative keywords such as '*보다* ([*bo-da*]: than)', '*가장* ([*ga-jang*]: most)', '*다르* ([*da-reu*]: different)', '*같* ([*gat*]: same)'. But many sentences also express comparison without those keywords. Similarly, although some sentences contain some keywords, they cannot be comparative sentences. By these reasons, extracting comparative sentences is not a simple or easy problem. It needs more complicated and challenging processes than only searching out some keywords for extracting comparative sentences.

Jindal and Liu (2006) previously studied to identify English comparative sentences. But the mechanism of Korean as an agglutinative language and that of English as an inflecting language have seriously different aspects. One of the greatest differences related to our work is that there are Part-of-Speech (POS) Tags for comparative and superlative in English[1], whereas, unfortunately, the POS tagger of Korean does not provide any comparative and superlative tags because the analysis of Korean comparative is much more difficult than that of English. The major challenge of our work is therefore to identify comparative sentences without comparative and superlative POS Tags.

We first survey previous studies about the Korean comparative syntax and collect the corpus of Korean comparative sentences from the Web. As we refer to previous studies and investigate real comparative sentences form the collected corpus, we can construct the set of comparative keywords and extract comparative-sentence candidates; the sentences which contain one or more element of the keyword set are called comparative-sentence candidates. Then we use some machine learning techniques to eliminate non-comparative sentences from those candidates. The final experimental results in 5-fold cross

[1] JJR: adjective and comparative, JJS: adjective and superlative, RBR: adverb and comparative, and RBS: adverb and superlative

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 153–156,
Suntec, Singapore, 4 August 2009.

validation show the overall precision of 88.68% and the overall recall of 88.40%.

The remainder of the paper is organized as follows. Section 2 describes the related work. In section 3, we explain comparative keywords and comparative-sentence candidates. In section 4, we describe how to eliminate non-comparative sentences from the candidates extracted in preceding section. Section 5 presents the experimental results. Finally, we discuss conclusions and future work in section 6

2 Related Work

We have not found any direct work on automatically extracting Korean comparative sentences. There is only one study by Jindal and Liu (2006) that is related to English. They used comparative and superlative POS tags and additional some keywords to search English comparative sentences. Then they used Class Sequential Rules and Naïve Bayesian learning method. Their experiment showed a precision of 79% and recall of 81%.

Our research is closely related to linguistics. Ha (1999) described Korean comparative constructions with a linguistic view. Oh (2003) discussed the gradability of comparatives. Jeong (2000) classified the adjective superlative by the type of measures.

Opinion mining is also related to our work. Many comparative sentences also contain the speaker's opinions and especially comparison is one of the most powerful tools for evaluation. We have surveyed many studies about opinion mining (Lee et al., 2008; Kim and Hovy, 2006; Wilson and Wiebe, 2003; Riloff and Wiebe, 2003; Esuli and Sebastiani, 2006).

Maximum Entropy Model is used in our technique. Berger et al. (1996) described Maximum Entropy approach to National Language Processing. In our experiments, we used Zhang's Maximum Entropy Model Toolkit (2004). Naïve Bayesian classifier is used to prove the performance of MEM (McCallum and Nigam (1998)).

3 Extracting Comparative-sentence Candidates

In this section, we define comparative keywords and extract comparative-sentence candidates by using those keywords.

3.1 Comparative keyword

First of all, we classify comparative sentences into six types and then we extract single comparative keywords from each type as follows:

Table 1. The six types of comparative sentences

	Type	Single-keyword Examples
1	Equality	'*같* ([*gat*]: same)'
2	Similarity	'*비슷하* ([*bi-seut-ha*]: similar)'
3	Difference	'*다르* ([*da-reu*]: different)'
4	Greater or lesser	'*보다* ([*bo-da*]: than)'
5	Superlative	'*가장* ([*ga-jang*]: most)'
6	Predicative	No single-keywords

We can easily find such keywords from the various sentences in first five types, while we cannot find any single keyword in the sentences of type 6.

Ex1) *"X 껌의 원재료는 초산비닐수지인데, Y 껌은 천연치클이다."* ([*X-gum-eui won-jae-ryo-neun cho-san-vi-nil-su-ji-in-de, Y-gum-eun cheon-yeon-chi-kl-i-da*]: Raw material of gum X is polyvinyl acetate, but that of Y is natural chicle.)[2]

And we can find many non-comparative sentences which contain some keywords. The following example (Ex2) shows non-comparative though it contains '*같* ([*gat*]: It means 'same', but it sometimes means 'think')'.

Ex2) *"내 생각엔 내일 비가 올 것 **같**아요."* ([*Nae sang-gak-en nae-il bi-ga ol geot gat-a-yo*]: I think it will rain tomorrow.)

Thus all the sentences can be divided into four categories as follows:

Table 2. The four categories of the sentences

Single-keyword	Contain	Not contain
Comparative Sentences	S1	S2
Non-comparative Sentences	S3	S4 (unconcerned group)

[2] In fact, type 6 can be sorted as non-comparative from linguistic view. But the speaker is probably saying that Y is better than X. This is very important comparative data as an opinion. Therefore, we also regard the sentences containing implicit comparison as comparative sentences

Our final goal is to find an effective method to extract S1 and S2, but single-keyword searching just outputs S1 and S3. In order to capture S2, we added long-distance-words sequences to the set of single-keywords. For example, we could extract '<는 [*neun*], 인데 [*in-de*], 은 [*eun*], 이다 [*i-da*]>' as a long-distance-words sequence from Ex1-sentence. It means that the sentence is formed as < S V but S V> in English (S: subject phrase, V: verb phrase). Thus we defined comparative keyword in this paper as follows:

Definition (comparative keyword)*: A comparative keyword is formed as a word or a phrase or a long-distance-words sequence. When a comparative keyword is contained in any sentence, the sentence is most likely to be a comparative sentence. (We will use an abbreviation 'CK'.)*

3.2 Comparative-sentence Candidates

We finally set up a total of 177 CKs by human efforts. In the previous work, Jindal and Liu (2006) defined 83 keywords and key phrases including comparative or superlative POS tags in English; they did not use any long-distance-words sequence.

Keyword searching process can detect most of comparative sentences (S1, S2 and S3)[3] from original text documents. That is, the recall is high but the precision is low. We here defined a comparative-sentence candidate as a sentence which contains one or more elements of the set of CKs. Now we need to eliminate the incorrect sentences (S3) from those captured sentences. First, we divided the set of CKs into two subsets denoted by CKL1 and CKL2 according to the precision of each keyword; we used 90% of the precision as a threshold value. The average precision of comparative-sentence candidates with a CKL1 keyword is 97.44% and they do not require any additional process. But that of comparative-sentence candidates with a CKL2 keyword is 29.34% and we decide to eliminate non-comparative sentences only from comparative sentence candidates with a CKL2 keyword.

4 Eliminating Non-comparative Sentences from the Candidates

To effectively eliminate non-comparative sentences from comparative sentence candidates with a CKL2 keyword, we employ machine learning techniques (MEM and Naïve Bayes). For feature extraction from each comparative-sentence candidate, we use continuous words sequence within the radius of 3 (the window size of 7) of each keyword in the sentence; we experimented with radius options of 2, 3, and 4 and we achieved the best performance in the radius of 3. After determining the radius, we replace each word with its POS tag; in order to reflect various expressions of each sentence, POS tags are more proper than lexical information of actual words. However, since CKs play the most important role to discriminate comparative sentences, they are represented as a combination of their actual keyword and POS tag. Thus our feature is formed as "$X \rightarrow y$". ('X' means a sequence and 'y' means a class; y_1 denotes comparative and y_2 denotes non-comparative). For instance, '*<pv etm nbn 기/pa ep ef sf >*[4] $\rightarrow y_2$' is one of the features from the sentence of *Ex2* in section 3.1.

5 Experimental Results

Three trained human annotators compiled a corpus of 277 online documents from various domains. They discussed their disagreements and they finally annotated 7,384 sentences. Table 3 shows the number of comparative sentences and non-comparative sentences in our corpus.

Table 3. The numbers of annotated sentences

Total	Comparative	Non-comparative
7,384	2,383 (32%)	5,001 (68%)

Before evaluating our proposed method, we conducted some experiments by machine learning techniques with all the unigrams of total actual words as baseline systems; they do not use any CKs. The precision, recall and F1-score of the baseline systems are shown at Table 4.

Table 4. The results of baseline systems (%)

Baseline System	Precision	Recall	F1-score
NB	35.98	91.62	51.66
MEM	78.17	63.34	69.94

The final overall results using the 5-fold cross validation are shown in Table 5 and Figure 1.

[3] As you can see in the experiment section, keyword searching captures 95.96% comparative sentences.

[4] The labels such as '*pv*', '*etm*', '*nbn*', etc. are Korean POS Tags

Table 5. The results of our proposed method (%)

Method	Precision	Recall	F1-score
CKs only	68.39	95.96	79.87
CKs + NB	85.42	88.59	86.67
CKs + MEM	***88.68***	***88.40***	***88.54***

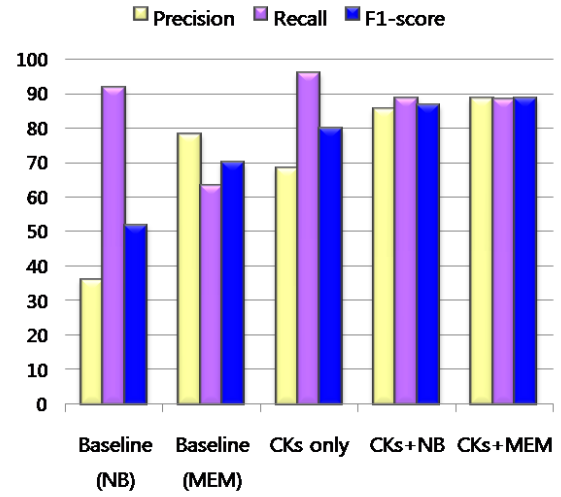

Fig. 1 The results of our proposed method (%)

As shown in Table 5 and Figure 1, both of MEM and NB is shown good performance but the F1-score of MEM is little higher than that of NB. By applying machine learning technique to our method, we can achieve high precision while we can preserve high recall.

6 Conclusions and Future Work

In this paper, we have presented how to extract comparative sentences from Korean text documents by keyword searching process and machine learning techniques. Our experimental results showed that our proposed method can be effectively used to identify comparative sentences. Since the research of comparison mining is currently in the beginning step in the world, our proposed techniques can contribute much to text mining and opinion mining research.

In our future work, we plan to classify comparative types and to extract comparative relations from identified comparative sentences.

Acknowledgement

This paper was supported by the Korean Research Foundation Grant funded by the Korean Government (KRF-2008-331-D00553)

References

Adam L. Berger et al. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39-71.

Andrea Esuli and Fabrizio Sebastiani. 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining. *European Chapter of the Association for Computational Linguistics*, 193-200.

Andrew McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naïve Bayes Text Classification. *Association for Advancement of Artificial Intelligence*, 41-48.

Dong-joo Lee et al. 2008. Opinion Mining of Customer Feedback Data on the Web. *International Conference on Ubiquitous Information Management and Community*, 247-252.

Ellen Riloff and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. *Empirical Methods in Natural Language Processing.*

Gil-jong Ha. 1999. *Korean Modern Comparative Syntax*, Pijbook Press, Seoul, Korea.

Gil-jong Ha. 1999. Research on Korean Equality Comparative Syntax, *Association for Korean Linguistics*, 5:229-265.

In-su Jeong. 2000. Research on Korean Adjective Superlative Comparative Syntax. *Korean Han-min-jok Eo-mun-hak*, 36:61-86.

Kyeong-sook Oh. 2004. The Difference between 'Man-kum' Comparative and 'Cheo-rum' Comparative. *Society of Korean Semantics*, 14:197-221.

Nitin Jindal and Bing Liu. 2006. Identifying Comparative Sentences in Text Documents, *Association for Computing Machinery/Special Interest Group on Information Retrieval*, 244-251.

Nitin Jindal and Bing Liu. 2006. Mining Comparative Sentences and Relations, *Association for Advancement of Artificial Intelligence*, 1331-1336.

Soomin Kim and Eduard Hovy. 2006. Automatic Detection of Opinion Bearing Words and Sentences. *Computational Linguistics/Association for Computational Linguistics.*

Theresa Wilson and Janyce Wiebe. 2003. Annotating Opinions in the World Press. *Special Interest Group in Discourse and Dialoque/Association for Computational Linguistics.*

Zhang Le. 2004. *Maximum Entropy Modeling Toolkit for Python and C++*. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.