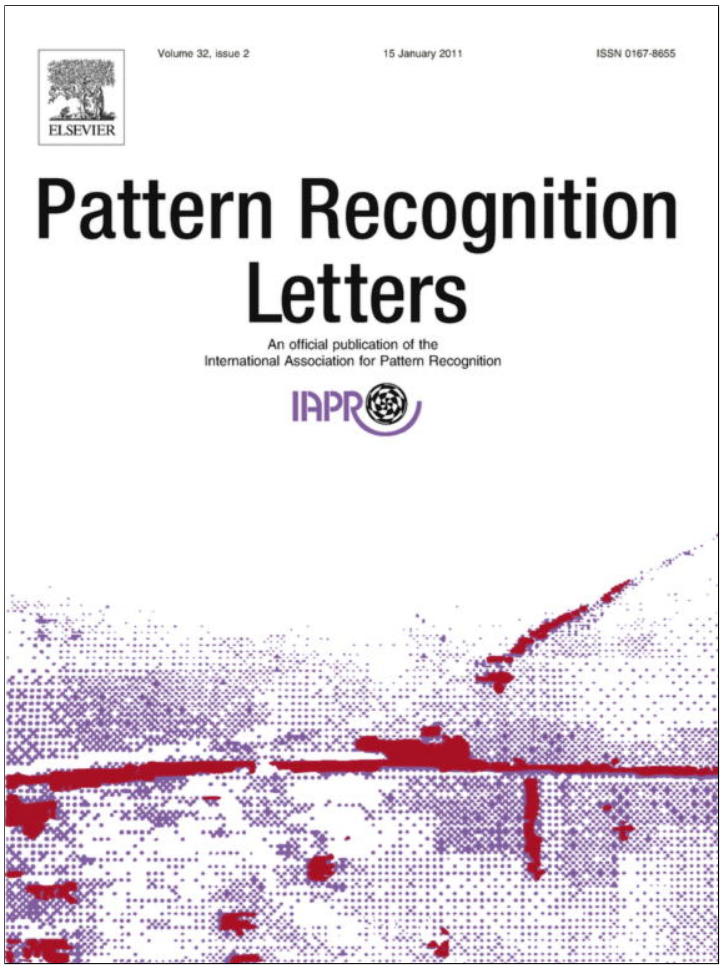

Pattern Recognition Letters 32 (2011) 293–296

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Finding relevant features for Korean comparative sentence extraction

Seon Yang, Youngjoong Ko *

*Dept. of Computer Engineering, Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan 604-714, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 28 January 2010
Available online 19 September 2010
Communicated by R.C. Guido

*Keywords:*
Comparative sentence
Comparative keyword
Comparative feature
Machine learning technique

ABSTRACT

In this paper, we study how to extract comparative sentences from Korean text documents. We decompose our task into three steps: (1) collecting comparative keywords; (2) extracting comparative-sentence candidates by keyword searching; and (3) eliminating non-comparative sentences from these candidates using machine learning techniques. We perform various experiments to find relevant features. As a result, our experiments show significant performance, an F1-score of 90.23%.



## 1. Introduction

In many cases, people should choose only one of two (or three or more) objects. For example, if you were trying to decide which item to buy between an *iPhone* and *Galaxy-S*, you would probably look for information on the Web to assist with your choice. Typing those two items into a search engine such as Google, Yahoo, or Naver would allow you to successfully find documents related to them. You could then open and read each retrieved document until you found enough information to make an informed decision about which to buy. It is obvious that getting information from the Web is a much better solution than previous methods. However, it is also clear that reading each document would still be a time-consuming job. Therefore, a comparison mining system capable of providing a summary of a comparison between two entities would be very useful in many areas such as marketing.

In this paper, we study the problem of extracting comparative sentences in Korean text documents. Our final goal is to build a Korean comparison mining system. Three tasks are needed for that: (1) extracting comparative sentences from text documents; (2) classifying those sentences into several classes; (3) analyzing comparative relations per each comparative class. Among these three tasks, this paper aims at the first task.

In previous work, Jindal and Liu (2006) studied the problem of identifying comparative sentences in English texts. 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, unlike English, the Korean Part-of-Speech (POS) Tagger does not provide the comparative and the superlative tags. Therefore, the major challenge of our work in such a language environment is, by finding relevant features, to identify comparative sentences with high performance enough for practical use.

Our task is composed of the following three steps: (1) collecting comparative keywords; (2) extracting comparative-sentence candidates by keyword searching; and (3) eliminating non-comparative sentences from the extracted candidates using machine learning techniques. As there is no previous study for Korean comparatives extraction directly related ours, we need to first collect texts from the Web and manually annotate them in order to generate a training/testing dataset.

After the annotation task, we investigate many real comparative sentences referring to studies on Korean Linguistics (Ha, 1999; Oh, 2004; Jeong, 2000), and then perform various experiments to find relevant features for our task. Finally, two types of features are found. One is the set of comparative keywords that are used to extract comparative-sentence candidates; a sentence that contains one or more elements of the keyword set is called a comparative-sentence candidate. The other is the set of sequential patterns that is used for machine learning techniques to eliminate non-comparative sentences from these candidates. The final experimental results in 5-fold cross validation show the overall precision of 92.24% and the overall recall of 88.31%.

The remainder of the paper is organized as follows. Section 2 describes the related work. Section 3 explains how to select comparative keywords as our first features and Section 4 explains how to extract sequential patterns as our second features. Section 5 presents the reason why comparative-sentence candidates are separated into two groups. Section 6 is devoted to the analysis of our experimental results. We finally conclude with a discussion of future work in Section 7.

* Corresponding author. Tel.: +82 51 200 7782; fax: +82 51 200 7783.
*E-mail addresses:* seony.yang@gmail.com (S. Yang), yjko@dau.ac.kr (Y. Ko).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.09.003

**Table 1**
The six types of comparative sentences.

| | Type | Example sentence | Single-CK |
|---|---|---|---|
| 1 | Equality | "X 와 Y 는 가격이 **같다**."<br>([X-wa Y-neun ga-gyeok-i gat-da]:<br>"X and Y are equal in price.") | '*같 ([gat]: same)*' |
| 2 | Similarity | "X 의 디자인은 Y 하고 **비슷하**네요."<br>([ X-ui di-ja-in-eun Y-ha-go bi-seut-ha-ne-yo]:<br>"The design of X is similar to that of Y.") | '*비슷하 ([bi-seut-ha]: similar)*' |
| 3 | Difference | "X 는 그 점에서 Y 와 **차이**가 있어요."<br>([ X-neun geu jeom-e-seo Y-wa cha-i-ga iss-eo-yo]:<br>"X differs from Y on that point.") | '*차이 ([cha-i]: difference)*' |
| 4 | Greater or lesser | "X 는 Y **보다** 성능이 뛰어나다."<br>([X-neun Y-bo-da seong-neung-i ddwi-eo-na-da]:<br>"X has better performance than Y.") | '*보다 ([bo-da]: than)*' |
| 5 | Superlative | "후보들 중에서 X 가 **가장** 신뢰가 간다."<br>([Hu-bo-deul jung-e-seo X-ga ga-jang sin-rwe-ga gan-da]:<br>"X is the most reliable among the candidates.") | '*가장 ([ga-jang]: most)*' |
| 6 | Implicit comparison | "X-바나나우유**는** 진짜 바나나로 만들**지만**, Y-바나나우유**는** 바나나 향으로만 맛을 낸**다**."<br>([X-ba-na-na-u-yu-neun jin-jja ba-na-na-ro man-deul-ji-man, Y-ba-na-na-u-yu-neun ba-na-na hyang-eu-ro-man mas-eul naen-da]: "Banana milk X contains real bananas, but banana milk Y is just banana aroma flavored.") | (No single-CK) |

## 2. Related work

Both linguistics and computer science are related to our research.

Although researchers in linguistics are not interested in computationally identifying of comparative sentences from a text document, they have focused on defining the syntax and semantics of comparative constructs. Ha (1999) described Korean comparative constructs. He classified comparative sentence structures into several classes and arranged comparison bearing words with a linguistic perspective. Oh (2004) discussed the gradability of comparatives and Jeong (2000) classified the adjective superlatives using certain measures.

As mentioned in the introduction section, we have found only one study done by Jindal and Liu (2006) for English comparative extraction. They used comparative and superlative POS tags, additional some keywords, class sequential rules, and the Naïve Bayesian learning method to search English comparative sentences. Their experiments showed the precision of 79% and the recall of 81% in English. There is no direct previous research on automatically extracting Korean comparative sentences until now.

As one research area of text mining, opinion mining is also related to our research because many comparative sentences also contain the speaker's opinion or sentiment. We have surveyed a lot of studies about opinion mining and sentiment classification (Lee et al., 2008; Kim and Hovy, 2006; Wilson and Wiebe, 2003; Riloff and Wiebe, 2003; Esuli and Sebastiani, 2006).

About machine learning techniques, we made use of the Maximum Entropy Model (Berger et al., 1996; Le, 2004), the Naïve Bayesian classifier (McCallum and Nigam, 1998), and Support Vector Machine (Joachims, 1998).

## 3. The first feature set: comparative keywords

In this section, we collect comparative keywords (hereafter, CKs) and then extract comparative-sentence candidates (hereafter, CS-candidates) using CKs. First of all, we classify comparative sentences into six types and then we extract single-CKs, which denote comparative keywords composed of only one word, from each type as follows:

We easily find single-CKs for the preceding five types in Table 1 from the various sentences while we cannot find any single-CK in the sentences of type 6. According to a linguistic point of view, the example sentence for type 6 in Table 1 can be sorted as a non-comparative sentence. But the speaker is probably saying that *X* is better than *Y*. Thus that kind of sentences can also be a very important data not only for comparison analysis but also for opinion/sentiment analysis. Therefore, we expand the scope of comparative sentences to include these implicit sentences.

The problem is that the example sentence for type 6 contains no single comparison-bearing word. However, we fortunately find out that a long-distance-words sequence '<*는 [neun], 지만 [ji-man], 는 [neun], 다 [da]*>'[1] can play a role of a CK for it. Hundred and seven long-distance-words sequences are added to the set of CKs and a total of 177 CKs are finally collected. Even though selecting these CKs is time consuming job, it is an only one time effort. Since keyword searching shows the recall of 95.96%, it is proven that our first feature set is successfully defined for CS-candidates extraction.

## 4. The second feature set: sequential patterns

In this section, we perform various experiments to eliminate non-comparative sentences from CS-candidates extracted in the previous section. Although the recall shown by keyword searching is sufficiently high, the precision of 68.39% shows that CKs also capture a lot of non-comparative sentences such as following sentence.

*"비가 올 것 같다."* (*[bi-ga ol geot gat-da*]: *I think it will rain.)*

The upper sentence is a non-comparative sentence that contains '*같 [gat]*' as one of CKs. '*같 [gat]*' means 'same' but sometimes has the meaning of 'conjecture'. In both cases, it has the adjective POS tag and we cannot distinguish these two cases with just POS tags. In order to classify these ambiguous cases, we employ machine learning techniques, the Naïve Bayesian classifier (NB), the Maximum Entropy Model (MEM) and Support Vector Machine (SVM). Our experimental process is as follows:

(1) Firstly, we conduct some experiments with all the unigrams (Case 1 in Table 2) and bigrams (Case 2) as our baseline systems; they do not use any CK.

[1] It means that the sentence is formed as ⟨S V but S V⟩ in English (S: subject phrase, V: verb phrase).

**Table 2**
Various features of the example sentence.

| | |
|---|---|
| Sentence | *"비가 올 것 같다." ([bi-ga ol geot gat-da]: I think it will rain.)* |
| Class | Non-comparative |
| CK | *'같 [gat]'* |
| POS tags[a] | *비/ncn 가/jcs 오/pv ㄹ/etm 것/nbn **같/pa** 다/ef ./sf* |
| Case 1: Unigram | *비, 가, 올, 것, 같, 다* |
| Case 2: Bigram | *비가, 가올, 올것, 것같, 같다* |
| Case 3: Lexical sequence (radius 1) | *<**같/pa**>, <것 같/pa>, <**같/pa** 다>, <것 **같/pa** 다>* |
| Case 4: POS tags sequence (radius 1) | *<**같/pa**>, <nbn **같/pa**>, <**같/pa** ef>, <nbn **같/pa** ef>* |
| Case 5: Combination | *< **같/pa** >, <nbn **같/pa** >, < **같/pa** 다>, <nbn **같/pa** 다>* |
| Case 6: POS tags sequence (radius 3) | *< **같/pa**>, <nbn **같/pa** >, < **같/pa** ef>, <nbn **같/pa**>, <etm nbn **같/pa**>, <nbn **같/pa** ef>, < **같/pa** ef sf>, <pv etm nbn **같/pa** >, <etm nbn **같/pa** ef>, <nbn **같/pa** ef sf>, <pv etm nbn **같/pa** ef>, <etm nbn **같/pa** ef sf>, <pv etm nbn **같/pa** ef sf>* |

[a] The labels such as *'ncn'*, *'jcs'* are Korean POS tags and *'ncn'* and *'nbn'* are noun POS tags.

**Table 3**
The numbers of annotated sentences.

| Total | Comparative | Non-comparative |
|---|---|---|
| 7384 | 2383 (32%) | 5001 (68%) |

(2) Secondly, we do an experiment with continuous lexical sequences within radius 1 of each CK (Case 3). As a result, this approach shows better performance than the baseline systems. Note that the features for sequence types (Cases 3, 4, 5, and 6) have the form of "$X \rightarrow y$" *('X' means a sequence and 'y' means a class; $y_1$ denotes comparative and $y_2$ denotes non-comparative).*
(3) In order to determine which one is more appropriate between the lexical sequence and the POS tags sequence, we compare the lexical sequence (Case 3) to the POS tags sequence (Case 4) and the combination of lexical and POS tags sequence (Case 5). In Case 5, we use POS tag for a word with a noun POS tag and we use lexical information for ones with the other POS tags.
(4) In addition, we conduct additional experiments with radius option of 2, 3, 4 and 5 for choosing the best radius option. As a result, the POS tags sequences with the radius option of 3 (Case 6) shows the best performance.
(5) As a result, we define our second feature set as follows:

– All the continuous sequences within the radius 3 of each CK.
– All the words except CK have just POS tag information.

## 5. Dividing CS-candidates into two groups

Since a sentence, which contains CKs such as an adverb '*보다* *([bo-da]*: than)', is almost likely to be the real comparative sentence, it needs no further process. Thus we divide our extracted CS-candidates into two groups, 'CKL1' and 'CKL2'. CKL1 includes the candidates that are retrieved by CKs showing more than 90% precision. The remained candidates are included in CKL2. The average precision of CKL1 is 97.44% while that of CKL2 is 29.34%. We finally decide to eliminate non-comparative sentences only from CKL2.

## 6. Experimental evaluation

Three trained human annotators compiled a corpus of 277 online documents from various domains. They discussed their disagreements and they finally annotated 7384 sentences. Table 3 shows the number of comparative sentences and non-comparative sentences in our corpus.

Table 4 compares the performance of the system with lexical sequences within the radius 1 of each CK to that of baseline systems. We achieved the best performance when we used the lexical sequences. As you can see in Table 4, SVM shows the best performance among three machine learning techniques: NB, MEM, and SVM. Therefore, all the performances in the following tables are reported as those of SVM.

Table 5 compares the performance of the system with the lexical sequences with that with POS tags sequences and that with the combination method. Experimental results using five radius options showed that the POS tags sequences are more relevant than the other two sequences and the radius option of 3 is most suitable among the five radius options. As a result, we determined the POS tags sequence within the radius 3 of each CK as our proposed feature. Fig. 1 shows the changes of F1-scores for the three sequences according to radius options.

**Table 4**
Unigram, bigram, and lexical sequences (%).

| Features | | Precision | Recall | F1-score |
|---|---|---|---|---|
| Baseline (unigrams) | NB | 35.98 | 91.62 | 51.66 |
| | MEM | 78.17 | 63.34 | 69.94 |
| | **SVM** | **87.86** | **72.57** | **79.49** |
| Baseline (bigrams) | NB | 34.43 | 96.13 | 50.70 |
| | MEM | 69.59 | 55.41 | 61.66 |
| | SVM | 80.15 | 68.26 | 73.73 |
| Lexical sequences | NB | 84.08 | 86.56 | 85.30 |
| | MEM | 87.14 | 86.24 | 86.69 |
| | **SVM** | **87.06** | **87.65** | **87.35** |

**Table 5**
Comparison of performances according to radius options (%); overall, the greater radius option increased the precision value but decreased the recall value. The best F1-score is shown when we use the POS tags sequences with the radius option of 3.

| Radius option | Lexical sequences | | *POS* tags sequences | | Combinations | |
|---|---|---|---|---|---|---|
| | Precision, recall | F1-score | Precision, recall | F1-score | Precision, recall | F1-score |
| 1 | (p) 87.06 (r) 87.65 | 87.35 | (p) 87.65 (r) 88.74 | 88.19 | (p) 86.85 (r) 88.02 | 87.43 |
| 2 | (p) 87.24 (r) 87.53 | 87.38 | (p) 90.47 (r) 88.56 | 89.50 | (p) 88.52 (r) 87.99 | 88.25 |
| 3 | (p) 88.02 (r) 87.21 | 87.61 | (p) 92.24 (r) 88.31 | *90.23* | (p) 89.47 (r) 87.87 | 88.66 |
| 4 | (p) 88.73 (r) 86.53 | 87.62 | (p) 93.48 (r) 87.09 | 90.17 | (p) 90.38 (r) 86.78 | 88.54 |
| 5 | (p) 90.11 (r) 85.14 | 87.55 | (p) 94.24 (r) 86.35 | 90.10 | (p) 90.76 (r) 86.20 | 88.42 |

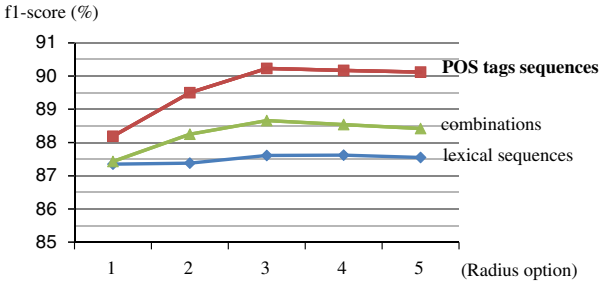


**Fig. 1.** F1-scores according to radius options; we determined the POS tags sequence with the radius of 3 as our proposed feature.

**Table 6**
Performance comparison between POS tags sequences and the other sequences with the radius of 3 (%); + indicates statistically significance at $p < 0.05$ and ++ indicates very statistically significance at $p < 0.01$.

| Features | Precision | Recall | F1-score |
|---|---|---|---|
| Lexical sequences | 88.02 | 87.21 | 87.61 |
| POS tags sequences | 92.24++ | 88.31+ | 90.23++ |
| Combinations | 89.47 | 87.87 | 88.66 |

**Table 7**
The division method for CKL1 and CKL2 (%).

| Target CS-candidates for machine learning | Precision | Recall | F1-score |
|---|---|---|---|
| CKL1 + CKL2 | 88.42 | 86.20 | 87.30 |
| CKL2 | *92.24* | *88.31* | *90.23* |

In order to check whether the difference between the proposed feature and the lexical sequences (or the combinations) were statistically significant, we performed a *t*-test at $p < 0.05$ level and $p < 0.01$ level. As given in Table 6, the proposed feature is statistically different in the precision and F1-score significantly.

Finally, we tried to prove that dividing CS-candidates into CKL1 and CKL2 in Section 5 is more effective than extracting comparative sentences from all the CS-candidates (CKL1 + CKL2). As shown in Table 7, the result confirmed that it is an improved method.

## 7. Conclusions

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

In our future work, we plan to complete Korean comparison mining system. First, we will classify comparative sentences into several classes, and then extract comparative relations from identified comparative sentences.

## Acknowledgment

This work was supported by the Dong-A University research fund.

## References

Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., 1996. A maximum entropy approach to natural language processing. Comput. Linguist. 22 (1), 39–71.
Esuli, A., Sebastiani, F., 2006. Determining term subjectivity and term orientation for opinion mining. In: European Chapter of the Association for Computational Linguistics, pp. 193–200.
Ha, G., 1999a. Korean Modern Comparative Syntax. Pijbook Press, Seoul, Korea.
Ha, G., 1999b. Research on Korean equality comparative syntax. Assoc. Korean Linguist. 5, 229–265.
Jeong, I., 2000. Research on Korean adjective superlative comparative syntax. Korean Han-min-jok Eo-mun-hak 36, 61–86.
Jindal, N., Liu, B., 2006a. Identifying comparative sentences in text documents. In: Association for Computing Machinery/Special Interest Group on Information Retrieval, pp. 244–251.
Jindal, N., Liu, B., 2006. Mining comparative sentences and relations. In: Association for Advancement of Artificial Intelligence, pp. 1331–1336.
Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning, pp. 137–142.
Kim, S., Hovy, E., 2006. Automatic detection of opinion bearing words and sentences. Comput. Linguist. Assoc. Comput. Linguist..
Le, Z., 2004. Maximum Entropy Modeling Toolkit for Python and C++. <http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html>.
Lee, D., Jeong, O., Lee, S., 2008. Opinion mining of customer feedback data on the Web. In: International Conference on Ubiquitous Information Management and Community, pp. 247–252.
McCallum, A., Nigam, K., 1998. A comparison of event models for Naïve Bayes text classification. Assoc. Adv. Artif. Intell., 41–48.
Oh, K., 2004. The difference between 'Man-kum' comparative and 'Cheo-rum' comparative. Soc. Korean Semantics 14, 197–221.
Riloff, E., Wiebe, J., 2003. Learning extraction patterns for subjective expressions. Empirical Methods Nat. Lang. Process..
Wilson, T., Wiebe, J., 2003. Annotating opinions in the world press. In: Special Interest Group in Discourse and Dialogue/Association for Computational Linguistics, pp. 11–12.