Journal of the Research Institute of Industrial Technology. Vol. 20, 2003

Extracting Representative Sentences for Summarization

*Jae-Hoon Kim * · Ho-Jin Park***

**Division of Mechinical and Information, National Korea Maritime University, Busan 606-791, Korea*

***Research Institute, Korea WISEnut, Inc.*

ABSTRACT : *This paper presents a system for extracting the representative sentences as part of the summary of a text. This system is called an indicative text summarizer To do this goal, we use some collocations as features for the sentence vectors. The collocations are likely to be compound nouns and syntactic relations and are automatically found using t-test as a kind of hypothesis test. Through some preliminary experiments, we have observed that the collocations of compound nouns and syntactic relations are very useful as features for Korean text summarization.*

KEY WORDS : aggregate similarity, compound nouns, collocation, Korean text summarization

1. Introduction

The World Wide Web is very large distributed digital information space called a cyberspace and has grown to encompass diverse information resources such as personal home pages, research publications, product and service catalogs, government announcements, and so on. Some existing search engines such as AltaVista[1], Yahoo[2], etc. are able to search and retrieve information from the Web efficiently and effectively, but retrieve too many documents, of which only small part are relevant to the user query. The technology of automatic text summarization is indispensable for dealing with this problem [1-3].

Text summarization is the process of extracting the most important information from a source (or sources) to get an abstract version for a particular user and goal [4]. In other words, text summarization is information compression. There are many uses of summarization in daily activity, for instance, headline news on TV and newspapers, minutes of several meetings, reviews of a book, and so on. Now, input to a summarizer could be not only a single document, but also multiple documents [5] and multimedia information such as image, audio, or video.

With the rapid growth of the Web and online information services, there has been an increase in the research and development funds devoted to this area and many researchers exert themselves to improve the performance of text summarization systems lately. Furthermore, the systems are appeared in several areas including commercial areas, which are in word processing tools (e.g. Microsoft's AutoSummarize) and in filters for web-based information retrieval (e.g. Inxight's LinguisticX[3] used in AltaVista Discovery).

Generally the process of automatic text summarization can be decomposed into three phases: analysis, transformation and synthesis [6]. We analyze a document into useful features like a list of nouns and collections, and a frequency of words in the analysis phase, transform it into a summary representation in the transformation phase, and generate a proper output form in order to read the summary easily in the synthesis phase.

A text summarization system can be broadly characterized as frequency-based, knowledge-based, or discourse-based [7]. These categories correspond to a continuum of increasing text understanding and increasing complexity in text processing. The frequency-based approaches rely on lexical and location information with the text, for instance, word-counts, proximity between words, locations, and cue phrases. Recently these approaches have used an automated method to combine these types of feature sets using training techniques [8-9]. Knowledge-based approaches depend on rich domain knowledge to interpret the conceptual structure of the text using techniques for natural language parsing, and can generate a highly readable summary. The discourse-based approaches are grounded in theories of text cohesion and coherence, and typically focus on linguistic processing of the text to identify the best cohesive sentence candidates. These approaches may combine with each other or heuristics to handle readability-related issues or to improve the performance.

There have been many research works on Korean text summarization [9-12]. They are likely to be immature and to be position on the spectrum of the frequency-based approaches. And they are difficult to compare with each other in performance objectively

* jhoon@mail hhu.ac kr 051)410-4574

** hanwool@wisenut co.kr

[1] http://www altavista.co.kr
[2] http://kr yahoo.com
[3] http //www.inxight.com

because we do not keep sharable test collections. So most of the systems had only evaluated in their own environments.

Salton *et al* (1999) represent a sentence by means of a graph called a text relationship map. In the graph, each node represents a vector of nouns in a sentence (or a paragraph), an undirected edge connects two nodes if two sentences (or two paragraphs) are semantically related. *i.e.* when the similarity between two sentences sufficiently large. The similarity is based on the word overlap between the corresponding sentences (or paragraph). A summary can be generated by extracting important sentences (or paragraphs) from text. Salton *et al* (1999) had measured the importance of a sentence (or a paragraph) as bushiness, of which a node representing a sentence (or a paragraph) on the map is defined as the number of edges connecting it to other nodes on the map. Since a highly bushy node is linked to many other nodes, it has an overlapping vocabulary with several nodes and is likely to discuss topics covered in many other nodes. Text summarization using the bushiness is to extract appropriate sentences (or paragraphs) with high bushiness and then to sort them in order of the bushiness.

In this paper, a text relationship map is redefined as a complete graph and the measure of the important sentence is also redefined as an aggregate similarity, which is a sum of weights on the edges instead of the number of edges connecting to other nodes on the map. We had already presented a Korean text summarization system using the aggregate similarity. In this paper, we improve the system by appending some collocations as features to the sentence vectors. The collocation is likely to be compound nouns and syntactic object relations and is extracted using *t* test.

The remainder of this paper is organized as follows: In Section 2, we describe sentence vectors constructed by features extracted from a Korean sentence. In Section 3, the computation of the aggregate similarity is described in detail. Then our proposed system for Korean text summarization is explicated in Section 4. We run experiments to show the usefulness of the collocations as features for Koran text summarization in Section 5 and compare with existing text summarization systems in Section 6. Finally, we draw conclusions with some discussions and future works.

2. Constructing Sentence Vectors

In this paper, a sentence in the text relationship map is represented by a feature vector of the form $S_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,n})$ where $s_{i,k}$ represents an importance value for feature N_k in sentence S_i. The importance values are computed by taking into account the frequency of features in a sentence. Accordingly, we first describe the method for extracting features from each sentence or document. Basically, the features only consisted of basic nouns [13] and, in this paper, are appended into some collocations, which is likely to be compound nouns and syntactic object relations

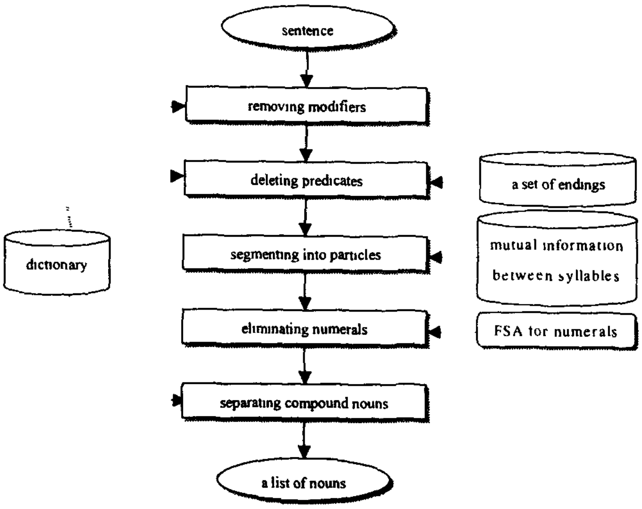

Figure 1. A system architecture for noun extraction and compound segmentation in Korean.

2.1. Extracting basic noun

Morphological analyzers are mainly used for noun extraction in Korean [14-15]. In this case, it is difficult as well as complex to implement the noun extraction system. Furthermore, it is not clear to resolve the ambiguity of morphological analysis and then part-of-speech tagging system is often used. To achieve the goal of speed and robustness, we do not use any linguistic tools like morphological analyzers and part-of-speech tagging systems. Figure 1 shows the system architecture for extracting basic nouns in Korean [16], of which the process is as the following steps:

1. removing modifiers with a non-nominal from a given sentence using a dictionary, where a modifier contains an adverb, an adnoun and an interjection;
2. deleting predicates from the remainder of the sentence using a dictionary and a list of endings, where a predicate is comprised of a verb and an adjective. A few of eojeols[4] carry some kinds of ambiguity between a predicate and a nominal. For example, typically eojeol "*narun*"[5]can be morphologically analyzed in three ways [17]: (1) "*na*/pronoun (I) + *run*/particle"[6], (2) "*na*/verb (spout) + *run*/ending", and (3) "*nal*/verb (fly) + *run*/ending." Here, the analysis (1) is a nominal and the analysis (2) and (3) are predicates. In this paper, such a ambiguity is ignored. In other words, we prefer the predicate to the nominal including a pronoun in case of showing this ambiguity. Because a pronoun is involved in a large number of nominals and we regard the pronoun as little import in this paper;

[4] In Korean, a word is called an eojeol, which is a sequence of morphemes surrounded by spaces. Hereafter, as for Korean, a word and an eojeol can be interchanged without notice

[5] In this paper, the Yale Romanization is used to represent Korean sentences and words.

[6] "*na*/pronoun (I)" means that the part-of-speech and the meaning of the morpheme '*na*' are pronoun and I, respectively.

3. segmenting a nominal into a noun and a particle using mutual information and some statistical information. To achieve this, we modify the Chinese word segmentation algorithm [18] and use the modified algorithm, which is not described in this paper in detail;
4. eliminating the numeral from the reminder of the sentence using the finite-state automata recognizing numerals. Here are the examples: eojeols "1999*nyen* (the year 1999)" and "10*kay*(10 pieces)";
5. separating a compound noun into basic nouns using a dictionary and the modified CYK parsing algorithm [19]. We choose longer nouns in preference to shorter nouns in length if the ambiguity is occurred in the process of the analysis.

2.2. Finding collocations

A collocation is an expression consisting of two or more words that corresponds to some conventional way of saying things [20]. In this paper, we are interested in two types of collections to improve the Korean text summarizer; one is a compound noun and the other is a syntactic object relation. To find these types of collocations, we use the hypothesis testing of *t* test [20]. To use the *t* test for finding collections, the *t* value is computed first as in Equation 1.

$$t = \frac{p(x,y) - p(x)p(y)}{\sqrt{\frac{p(x,y)}{N}}} \qquad (1)$$

In Equation 1, *x* and *y* represent Korean basic nouns in compound noun extraction and *x* is a object of a predicate *y* in syntactic relation extraction. *p*(*x*) is a probability occurring *x* in the corpus and *N* is total occurrences of a bigram *xy* in a document. If the *t* value is larger than a critical value, which is a kind of threshold value, we reject the null hypothesis[7] that is, *xy* is accepted as a collection. The critical value is controlled according to the reliability. In our experiments described afterwards, the reliability is very low as compared with the value introduced in [20] because the size of a document is not as large as the size of a corpus to find a collocation generally. In sequence, we describe the methods for finding collections as compound nouns and syntactic object relation. First, compound nouns as collocations are automatically extracted in the following steps:

1. obtaining a list of nouns in a documents through the method described in Section 2.1,
2. building a bigram of a list of nouns using a two word collocational window at a distance;
3. estimating *p(x)* for each noun *x* and *p(xy)* for each bigram *xy* by relative frequency;
4. finding collocations as compound nouns using *t* test described above;
5. appending each compound noun into the corresponding sentence vectors as features.

In finding syntactic object relations, we use two heuristics: one is that the object is the leftmost noun of a predicate (a verb or an adjective) in Korean; The other is that two different predicates is very similar semantically if first two letters of the two predicates is identical.[8] In this paper, we do not use one-letter predicates because they play a role in noise in feature vector and have too many ambiguities. Using these heuristics, we extract a collocation as syntactic object relation in the following steps:

1. obtaining a list of nouns and predicates in a documents using the method, which is modified from the step 2 of the method described in Section 2.1 to include predicates;
2. extracting pairs of leftmost nouns and predicates using the above heuristics;
3. estimating *p(x)* for each noun *x* or each predicate x, and *p(xy)* for each pair of noun and predicate *xy* by relative frequency;
4. finding collocations as compound nouns using *t* test described above
5. appending each compound noun into the corresponding sentence vectors as features.

3. Computation of Aggregate Similarity

Salton *et al.* (1999) depict a document in the graph called a text relationship map. In this paper, a text relationship map is redefined as a complete graph and the measure of the important sentence is also redefined as an aggregate similarity, which is a sum of weights on the edges instead of the number of edges connecting it to other nodes on the map. In the map, each node represents a sentence vector S_i as mentioned in Section 2, each edge connects two sentence vectors, S_i and S_j, and a weight on the edge is a value of the similarity between a pair of sentences. The vector similarity *sim*(*i*, *j*) can be computed as the inner product like Equation 2.

$$sim(i,j) = \sum_{k=1}^{n} s_{i,k} s_{j,k} \qquad (2)$$

where, *n* is the number of distinct features in the document, S_i is $(s_{i,1}, s_{i,2}, \ldots, s_{i,n})$, and $s_{i,k}$ is a frequency of feature N_k in sentence S_i. We measure the importance of sentence S_i as an aggregate

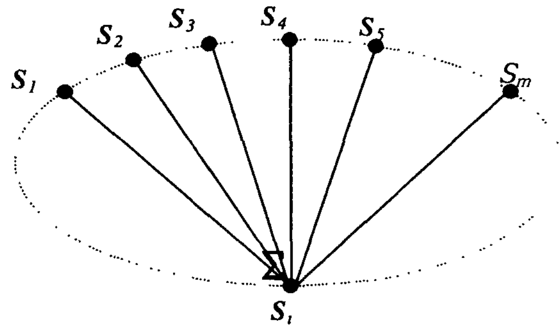

Figure 2. The concept diagram of an aggregate similarity.

[7] The null hypothesis is the occurrences of *x* and *y* are independent; H₀ : *p(x,y) = p(x)p(y)*,

[8] This heuristics may be wrong in some situations, but correct in many cases.

similarity *asim*(*i*), of which a node representing a sentence on the map is defined as the sum of weights on the edges connecting it to other nodes on the map like Equation 3. Since a highly important node is linked to many other nodes, it has an overlapping vocabulary with several nodes and is likely to discuss topics covered in many other nodes. In consequence a value of the aggregate similarity on the highly important node is relatively high in comparison with those on others. Figure 2 depicts the conceptual diagram of the aggregate similarity of a sentence S_i of *m* sentences in the document.

$$asim(i) = \sum_{j=1, j \neq i}^{n} sim(i, j) \qquad (3)$$

4. Korean Text Summarization System

This section describes in detail the process that generates a summary by selecting sentences based on the aggregate similarity as mentioned above. The value of aggregate similarity for each sentence is estimated based on occurrences of features like nouns, compound nouns, and syntactic object relations. The necessary statistical information about features is directly obtained from a given source document without any training process Accordingly our system is characterized by high adaptability and practicability. Figure 3 shows the distinct steps of our proposed system

The preprocessing phase segments a given source document into sentences based on punctuation marks ".!?:;" and removes some symbols like '(' and ')'. In this phase, some ambiguities can be occurred, for example, some word sequences such as "1999. 12."To resolve this problem, we use some heuristic rules case by case. The phases of the feature extraction and the computation of the aggregate similarity are clearly described in Section 2 and Section 3, respectively. The last phase of sentence selection rearranges sentences in the order of the value of the aggregate similarity, and then chooses as many sentences as you want.

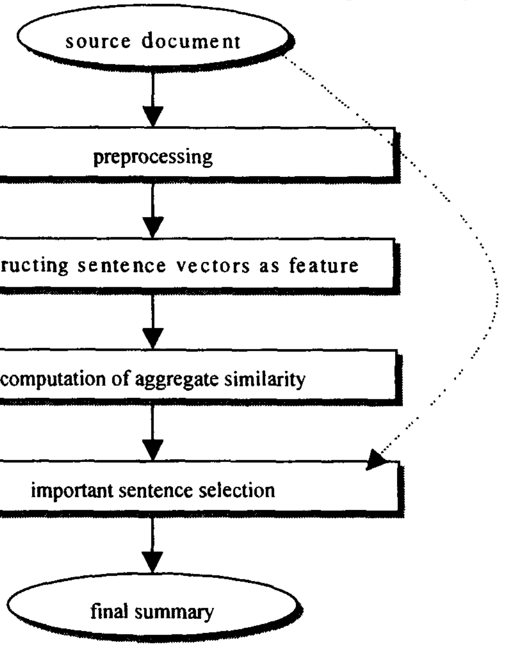
source document
preprocessing
constructing sentence vectors as feature
computation of aggregate similarity
important sentence selection
final summary

Figure 3. Overview of Korean text summarization system.

5. Evaluation

5.1. The Corpus

We used two test collections for evaluation: a paper collection (PAPER) and a newspaper collection (NEWS)[9] The paper collection consists of 100 articles with their author-written summaries sampled from the area of computer science. The documents are drawn from World Wide Web partially and CD-ROM distributed by Korea Information Science Society. We remove tables, equations, figures, captions, references, and cross references from the original documents and check them manually. In order to alignment between a summary and document sentences, we use an automatic method, assisted by the inner product similarity between author's abstract and sentences in the body (PAPER-ALL) or Introduction and Conclusion sections (PAPER-InCon) of the document. The number of sentences in the summary is the same as that in author's abstract. The newspaper collection is composed of 105 articles selected from KORDIC test collection [21]. All documents come from newspapers. Their summaries are created by human annotators. Table 1 shows the statistics of the two test collections. The test collections of PAPER-InCon and NEWS are very similar in statistics, but different from each other in genre.

Table 1. Statistics about test collections

test collection	PAPER ALL	PAPER-InCon	NEWS
the no. of documents	100.0	100.0	105.0
the average no. of sentences per document	113.2	20.6	21.5
the average no. of sentences per summary	5.6	5.6	6.0

5.2. Evaluation

Consider an example document for summarization. Let *A* be the set of sentences in its summary, *B* be the set of sentences extracted from a text summarization system, and *C* be the set of sentences in the intersection of the sets *A* and *B*. We employ measures of recall *R*, precision *P*, and F-measure *F*, which are widely used in information retrieval [21]. The definitions of the measures are as follows;

- Recall R is the proportion of the target sentences thatthe system extracted, *i.e.* $R = \frac{|C|}{|A|}$
- Precision P is the proportion of the extracted sentences that the

[9] available at ftp://nlplab.kmaritime ac.kr/pub/KoreanSummCollection/.

system got right, i.e. $P=\frac{|C|}{|B|}$

- F-measure F is to combine precision and recall into a single measure of overall performance, i.e $F=\frac{2PR}{P+R}$

Table 2 shows the best performance of our system according to the variations of the critical values for finding collocations described in next section.

The first column in Table 2 shows compression rates, which is the size of the summary as a fraction of the source document. The average number of sentences extracted from the test collection PAPER-ALL is relatively high in comparison with the other test collections because the number of sentences in the body of documents is very high. Consequently the performance of PAPER-ALL are appeared in contrary to that of the other test collections, especially PAPER. The characteristics of PAPER-InCon and NEWS are similar in statistics as mentioned above, but different from each other in performance. This is the reason as following. (1) it is different to construct PAPER-InCon and NEWS. The former is automatically built using author's abstract and the latter manually created by human's judge. PAPER-InCon outperforms NEWS. This is the reason why PAPER-InCon constructed automatically reflects its own idiosyncrasy, we think. (2) They are different from each other in genre. The expressions in technical documents like PAPER-InCon are regularly reiterated. However, we need to steadily observe this phenomenon in order to reveal clearer reasons.

5.3. Performance Improvement of Collocations

Table 3 shows the performance variation of compound nouns in proportion to the variation of critical values as model parameters. Collocations of compound nouns currently yield the best performance at the critical values of 1.645 (reliability of 90%) and 1.282 (reliability of 80%)[10] at the compress rate of 10% and 20%, respectively. The performance is not varied heavily according to the control of the critical value. The higher the compression rate is, the lower the reliability is. Compared with the values introduced in [20], the reliability is relatively low because the collections are extracted from a document, which is small in proportion to corpora in size of word

Table 4 shows the performance variation of syntactic object relations in proportion to the variation of critical values as model parameters. Collocations of syntactic object relations currently produce the best performance at the critical values of 1.960 (reliability of 95%) and 1.645 (reliability of 90%) at the compress rate of 10% and 20%, respectively. In the control of critical values, the result is very similar to that of compound nouns. Compared with the compound nouns, the reliability is highly controlled because the syntactic object relation is very likely to be noisy and error-prone.

Table 2. Performance of the proposed text summarization system.

compression rate	test collection	precision	recall	F-measure
10%	PAPER-ALL	33.5	58.8	42.7
	NEWS	44.4	18.0	23.6
	PAPER-InCon	83.2	33.0	23.6
20%	PAPER-ALL	20.5	72.3	31.9
	NEWS	43.5	27.1	33.4
	PAPER-InCon	76.6	46.2	57.6

Table 3. Performance variation according to critical value (compound nouns)

measure	compression rate	reliability (critical value)	PAPER-ALL	NEWS	PAPER-InCon
precision	10%	80% (1.282)	33.3	43.8	82.9
		90% (1.645)	**33.3**	**44.4**	**83.9**
		95%(1.960)	33.3	43.6	81.9
	20%	**80% (1.282)**	**20.5**	**43.5**	76.3
		90% (1.645)	20.4	43.3	**76.7**
		95%(1.960)	20.3	43.8	76.3
recall	10%	80% (1.282)	58.3	17.9	32.4
		90% (1.645)	**58.3**	**18.0**	**33.0**
		95%(1.960)	58.4	17.9	32.1
	20%	**80% (1.282)**	**72.3**	**27.1**	45.8
		90% (1.645)	72.2	26.8	**46.1**
		95%(1.960)	71.8	27.3	45.7

Table 5 shows the performance improvement by appending collocations into sentence vectors. "basic system" in the third column of Table 5 means that sentence vectors consist of nouns excluding compound nouns. "compound nouns" and "syntactic obj. relations" in third column of Table 5 means that sentence vectors are composed of nouns plus compound nouns and syntactic object relations, respectively. "+CN +SR"means that the sentence vectors are made of nouns, compound nouns, and syntactic object relations. The collocations of compound nouns and syntactic object relations are useful, but not dramatic. Our system has improved in the F-measure by about 1% in average since the extracted collocations are not sufficiently large in number and often unreliable in case of syntactic object relations. We, however, take much interest in collocations because the performance does not go down.

5.4. Korean Text Summarization Systems

In this section, we describe the characteristics of existing text summarization systems for Korean and make a comparison between our system and the existing systems. It is very difficult to make the comparison objectively because of their various experimental environments. Table 5 summarizes the features of the existing systems

[10] See the appendix of "tiny statistical tables" in [20] for more information on the relationship of the critical value and the reliability in t test.

and our system. We draw up the table based on their publications. Most of the existing systems are based on a statistical approach. And they had been evaluated for Introduction and Conclusion sections of papers on a very small scale. By contrast, we evaluate our system for a large size of articles in the newspapers as well as documents in technical documents.

Most of the existing systems need some training processes. The processes can automatically estimate model parameters without great labors, but require a large scale of training data, which is manually constructed with a great effort and time. Furthermore, the system with training process is hard to adapt to new environments. Our system do not require any training processes except noun extraction.

6. Conclusion

We had presented a Korean text summarization system using an aggregate similarity. The aggregate similarity means the sum of values of the similarities between sentences. Each similarity value is estimated using inner product. Our system is characterized by easy implementation, low cost of computation, easy adaptation, and practical use. In this paper, we improve the Koran text summarization system using collocations of compound nouns and syntactic object relations. The collocations are extracted from a document using *t* test.

To evaluate our system, we use two test collections: one collection (PAPER-InCon and PAPER-ALL) consists of 100 papers in the domain of computer science; the other collection (NEWS) is made of 105 articles in the newspapers. Under the compression rate of 20%, we achieved recall of 46.2% (PAPER-InCon) and 27.1% (NEWS), and precision of 76.6% (PAPER-InCon) and 43.5% (NEWS). We observed that the collections are useful through our experiments.

Since the research in this area, is at its infant stage in Korean, there

Table 4. Performance variation according to critical value (synactic object relations)

measure	compression rate	reliability (critical value)	PAPER-ALL	NEWS	PAPER-InCon
precision	10%	80%(1.282)	32.6	42.4	83.2
		90% (1.645)	32.7	42.4	83.2
		95%(1.960)	32.8	42.4	83.2
	20%	**80% (1.282)**	20.4	41.9	76.4
		90%(1.645)	20.4	42.0	76.4
		95%(1.960)	20.4	42.0	76.3
recall	10%	80%(1.282)	56.7	17.3	32.5
		90% (1.645)	56.7	17.3	32.5
		95%(1.960)	57.0	17.3	32.5
	20%	**80% (1.282)**	71.8	26.2	45.8
		90%(1.645)	71.8	26.2	45.8
		95%(1.960)	71.8	26.2	45.8

Table 5. Performance improvement based on collections

compression rate	test collection	sentence vector	precision	recall	F-measure
10%	PAPER-ALL	basic system	32.2	57.0	41.2
		+ compound nouns (CN)	33.3	58.3	42.4
		+ syntactic obj. relations (SR)	32.8	57.0	41.6
		+ CN + SR	**33.5**	**58.8**	**42.7**
	NEWS	basic system	42.4	17.9	25.2
		+ compound nouns (CN)	44.4	18.0	25.6
		+ syntactic obj. relations (SR)	42.4	17.3	24.6
		+ CN + SR	**44.4**	**18.0**	**25.6**
	PAPER-InCon	basic system	83.2	32.5	46.7
		+ compound nouns (CN)	83.9	33.0	47.4
		+ syntactic obj. relations (SR)	83.2	32.5	46.7
		+ CN + SR	**83.9**	**33.0**	**47.4**
20%	PAPER-ALL	basic system	20.4	71.8	31.8
		+ compound nouns (CN)	**20.5**	**72.3**	**31.9**
		+ syntactic obj. relations (SR)	20.4	71.8	31.8
		+ CN + SR	20.4	72.1	31.8
	NEWS	basic system	42.0	26.2	32.3
		+ compound nouns (CN)	43.5	27.1	33.4
		+ syntactic obj. relations (SR)	42.0	26.2	32.3
		+ CN + SR	**43.5**	**27.1**	**33.4**
	PAPER-InCon	basic system	76.6	46.0	57.5
		+ compound nouns (CN)	76.3	45.8	57.2
		+ syntactic obj. relations (SR)	76.4	45.8	57.3
		+ CN + SR	**76.6**	**46.2**	**57.6**

Table 6. Characteristics of Korean text summarization systems

(Myaeng and Jang) [8] 1999	- statistical and heuristic approach - training - sentence selection	- Introduction and Conclusion sections of papers (30)	53.19	39.53	5 sentences regardless of the size of source document
(Kang) [9] 1997	- statistical approach - training - sentence similarity - cosine - sentence selection	- Introduction and Conclusion sections of papers (25)	51.08	42.4	20%
(Lee *et al.*) [10] 1999	- statistical and heuristic approach - training - sentence similarity - sentence selection	- Introduction and Conclusion sections of papers(20)	66.8.0		30%
(Ryu and Lee) [11] 2000	- statistical approach - bushy path - cosine - paragraph selection	- Introduction and Conclusion sections of papers(25)		35.0	30%
Our system	- statistical approach - aggregate similarity - inner product - sentence selection	- Introduction and Conclusion sections of papers (100) - Articles in newspaper(105)	46.2 27.1	76.6 43.5	20%

are many things to be investigated in the future. First most of summaries generated by summanzers are hard to understand, especially based on a statistical approach. To overcome this, we need to apply to this area many techniques of Korean language processing like text planning and sentence generation. Second, the precision and the recall are not enough to use the summanzers practically. We require the improved systems in performance. To do this, we should improve the robust text analysis including robust anaphora resolution and topic detection and tracking. Third, we develop new linguistically annotated summarization corpora in different genres containing pairs of a document and the corresponding summary. These can help develop better comparison and evaluation of corpus-based methods as in the field of information retrieval and natural language processing. Finally we should improve availability of meta-information related to non-textual media such as multimedia and exploit advances in related fields of information visualization and information retrieval.

References

[1] Cowie, J., Mahesh, K., Nirenburg, S. and Zajac, R., "MINDS multilingual interactive document summarization", In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, pp. 131-132, 1998.

[2] Jang. D. and Myaeng, S.-H., "Automatic text summarization systems", *Korea Information Science Society Review*, vol. 15, no. 10, pp.42-49, 1997.

[3] Salton, G., Singhal, A., Mitra, M. and Buckly, C., "Automatic text structuring and summarization", In *Advances in Automatic Text Summarization*, Eds. Mani, I. and Maybury, M. T., The MIT Press, pp. 61-70, 1999.

[4] Mani, I. and Maybury, M. T., *Advanced in Automatic Text Summarization*, The MIT Press, 1999.

[5] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., and Eskin, E., "Towards multi-document summarization by reformulation : Progress and Prospect," In *Proceedings of AAAI-99*, 453-460, 1999.

[6] Sparck Jones, K., "Automatic summarizing: factors and directions", In *Advances in Automatic Text Summarization*, Eds Mani, I. and Maybury, M. T., The MIT Press, pp. 1-12, 1999.

[7] Aone, C., Gorlinsky, J., Larsen, B., and Okurowski, M. E., "A trainable summarizer with knowledge acquired from robust NLP techniques", In *Advances in Automatic Text Summarization*, Eds. Mani, I. and Maybury, M. T. The MIT Press, pp. 71-80, 1999.

[8] Kupiec, J., Pedersen, J., and Chen, F., "A trainable document summarizer", In *Proceedings of the 18th ACM-SIGIR Conference*, pp. 68-73, 1995.

[9] Myaeng, S. H. and Jang, D., "Development and evaluation of a statistically based document summarization system," In *Advances in Automatic Text Summarization*, Eds. Mani, I. and Maybury, M. T. The MIT Press, pp. 61-70, 1999.

[10] Kang, S.-B., *Implementation of a summarization system using statistical information of Korean documents*, Master's thesis, Pusan National University, Department of Computer Science,

1997.

[11] Lee, M.-H., Park, M.-S.. Kim, M.-J., and Lee, S.-J., "Sentence extraction using document features and heading," In *Proceedıngs of KIPS,*. vol. 6, no. 2, pp. AI41-AI45, 1997.

[12] Ryu, D.-W. and J.-H. Lee.,"Word co-occurrence based automatic text summarization", In *Proceedings of KISS,* vol. 27, no. 1, pp. 345-347, 2000.

[13] Kim, J.-H, Kim, J.-H. and Hwang, D-S. "Korean text summarization using an aggregate similarity", In *Proceedings of the 5thInternational Workshop on Information Retrıeval with Asian Languages,* Hong Kong, pp. 111-118, 2000.

[14] Kim, Y.-K. and Kwon, H.-C., "Noun extraction system in information retrieval system of MIRINE", In *Proceedings the 1st Workshop on the Evaluation for Morphological Analyzer and Part-of-Speech Tagging System,* pp. 89-91, 1999.

[15] Won, H., Park, M. and Lee. G., "Integrated indexing method using compound noun segmentation and noun phrase synthesis", *Journal of KISS. Software and Applicatıons,* vol 27, no. 1, pp. 84-95, 2000.

[16] Kim, J.-H., Kim, J.-H, and Park, H.-J., "Korean noun extraction with filtering and segmentation", in *Proceedings of the 1st International Conference on East-Asıan Language Processing and Internet Information Technology (EALPIIT2000),* Northeastern University, Shenyang, China, pp. 107-112, 2000.

[17] Kim, J.-H., "Korean part-of-speech tagging using a weighted network", *Journal of KISS (B) · Software and Applications,* vol. 25, no. 6, pp. 951-959, 1998.

[18] Maosong, S., Dayang, S. and Tsou, B. K., "Chinese word segmentation without using lexicon and hand-crafted training data", In *Proceedings of COLING-ACL* 98, pp. 1265-1271, 1998.

[19] Aho, A. V. and Ullman, J. D., *The Theory of Parsing, Translation, and Compiling,* Prentice-Hall, 1973.

[20] Manning, C. D. and Schütze, H., *Foundations of Statistical Natural Language Processing,* The MIT Press, 1999.

[21] Kim, T.-H., Park, H.-R., Shin, J.-.H., "A study on text understanding model for retrieval / summarization / filtering", In *Proceedings of the Workshop on Softscience,* 1999.

[22] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.

원고접수일 : 2002년 x월 x일
원고채택일 : 2003년 x월 x일