J Inf Process Syst, Vol.16, No.2, pp.447~459, April 2020
https://doi.org/10.3745/JIPS.02.0131

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Conceptual Extraction of Compound Korean Keywords

Samuel Sangkon Lee*

Abstract

After reading a document, people construct a concept about the information they consumed and merge multiple words to set up keywords that represent the material. With that in mind, this study suggests a smarter and more efficient keyword extraction method wherein scholarly journals are used as the basis for the establishment of production rules based on a concept information of words appearing in a document in a way in which author-provided keywords are functional although they do not appear in the body of the document. This study presents a new way to determine the importance of each keyword, excluding non-relevant keywords. To identify the validity of extracted keywords, titles and abstracts of journals about natural language and auditory language were collected for analysis. The comparison of author-provided keywords with the keyword results of the developed system showed that the developed system was highly useful, with an accuracy rate as good as up to 96%.

Keywords

Concept Word with Co-occurrence, Importance of the Keyword Candidate, Keyword Extraction, Keyword Pattern, Production Rule, Relation of Sentential Distance and Conceptual Distance

1. Introduction

Keywords are usually extracted from a document by selecting a few critical terms that arise from the text. Such a method of keyword extraction can be widely applied to different information retrieval technologies such as keyword indexing [1-14] and information extraction [5,10,11,13]. Automatic keyword extraction can be realized by analyzing information such as word frequency count and word placement [2,5,9,13,14] or by using natural language processing that looks more closely at sentence structure and word context [3,4,6,13].

However, if there are no words to serve as keywords, if some words composed of keywords are spread across the text body, or when the context of the text is to be guessed, term draw was used as a meaning of abstraction, derivation, or conjecture. This process is an inferring of the context, concept, or conclusion, not a detection of words in the document, so term draw is not effective when there are abstract vocabularies or keywords by which the system cannot infer or guess. Furthermore, this is not effective when there are no terms within the document that can be used as a valid keyword, when the potential keywords appear in too many places in the document, or when the keywords are made of too abstract terms that cannot be associated with the subject of the document (the word “extraction” was used in [14] in the context of derivation and implication).

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Manuscript received August 26, 2016; first revision December 6, 2016; accepted January 7, 2017.
Corresponding Author: Samuel Sangkon Lee (samuel@jj.ac.kr)
* Dept. of Computer Science and Engineering, Jeonju University, Jeonju, Korea (samuel@jj.ac.kr)

Copyright© 2020 KIPS

Concerning this issue, Nagata and Kimoto [6] defined an index rule explaining the basic concept of how to make up keywords (key concept) and describing concurrence (word sharing) among keywords. This rule served as their basis in their proposed keyword extraction method. In this rule, nouns are first detected from sentences, and a proper keyword is extracted by collecting words similar to or same with the nouns. An appropriate keyword to the text may not be analyzed as relevance among similar or same words is not considered when the key concept is extracted.

Against this backdrop, this study experiments on extracting new keywords by considering the relevance between and among concepts of the main words appearing in the document. To extract proper keywords as best as possible even if acceptable ones do not appear in the document itself, this study proposes a concept-based compound keyword extraction method. Particularly, when author-provided keywords do not appear in the text, or they are not selected as thematic words, a compound-word-production rule is used. Focused on the conceptual relevance of the production rule, a new calculation method to determine importance is introduced.

This study is structured as follows. Section 2 briefly analyzes author-provided keywords and provides an overview of a production rule setup. Section 3 proposes a concept-based production rule and the calculation method of keyword importance, considering the co-occurrence of keywords. Section 4 experiments for the validity of author-provided keywords and evaluates extracted compound keywords. Finally, Section 5 finishes with the conclusion and suggestions for future studies.

Table 1. Six keyword patterns

Case no.	Example sentence	Extracted pattern	Remarks
1	Speak a language and recognize it.	Language recognition	Extraction via demonstrative noun
2	It aims to process the human voice with a processor after successful voice recognition.	Speech recognition	Extraction using words appearing in multiple sentences
3	Word mining	Keyword extraction	Extraction via compound keyword transformation (use of thesaurus)
4	Humans have wanted to process a language in machines. Efforts for such have been made for tens of years.	Speech recognition	Analyzes concurrence of words existing in multiple sentences
5	Inferred knowledge Can attribute part of speech	Artificial intelligence Morphological analysis	Analyzes a relevant field and abstract vocabularies
6	Back-off Context-free grammar	Back-off Context-free grammar	Extracts by conversion of English or acronym

2. Keyword Patterns

There are some characteristics of keyword patterns, as seen in Table 1 selected and written by authors to extract correct keywords sticking to the theme of the document. The table displays six real examples (sentences or character strings) and extraction patterns. Each different pattern is adopted for each different

occasion, and the occasion falls into three categories. First, when keywords exist in all parts of the document. Second, when keywords exist in a part of the document. Third and last, when keywords do not exist in the document. The extraction patterns described above can be correctly analyzed using each concept word. As a pre-step to generate keywords based on the rule, compound keywords are divided into each component, and pattern analysis will be carried out. For case # 6 found in Table 1, a dictionary for the conversion should be generated. The next section will introduce the concept-based keyword extraction method.

3. Keyword Extraction

This section attempts to extract keywords based on concepts. Section 3.1 discusses production rule for forming compound keywords, and Section 3.2 discusses conceptual distance, the number of co-occurrences, and the calculation of keyword importance for ranking keyword candidates.

3.1 Production Rule for Compound Keywords

Compound keywords are not often found in texts requiring a guess or inference from words appearing in the texts. This section employs a method that Nagata and Kimoto [6] used to define relations between the concept of morpheme in keywords and the keywords themselves as a rule. Under the rule, it detects words embracing abstract vocabularies or themes [9].

Under the assumption that a compound keyword *w* is a keyword and the compound word consists of component morphemes, which are composed of sub-concept words, the component morphemes are called conceptual components of *w*. These conceptual components consist of synonyms and related terms of *w*.

The compound word, *w*, can be made up again of each component's synonyms and related terms. It can be structured as follows using [ ]. The first { } represents synonyms, and the second { }, related terms. Each component of w is expressed as *w1*, *w2*, ..., *wn*, and the function to obtain the *w*'s concept is Concept ( ). The conceptual components of each phoneme are as follows:

- Concept (meaning) = [{definition, value, ...}, {context, ...}]
- Concept (processing) = [{dispensing, disposing, ...}, {solving, handling, ...}]
- Concept (vocal) = [{voice, ...}, {timbre, intonation, ...}]
- Concept (dialogue) = [{conversation, discussion, ...}, {talk, chat, ...}]

The production rule (PR) of *w1, w2, ... wn* is defined as the addition of all conceptual components or PR(*w1*, *w2*, ..., *wn*) = Concept(*w1*) + Concept(*w2*) + … + Concept(*wn*). This extracts the compound keyword *w* or the set of *w1, w2, ..., wn* only when all concept elements from Concept(*w1*) to Concept(*wn*) appear in the document. For example, there is a compound word of "meaning processing" in a document. Under the production rule, if the context components ("meaning" and "processing") and two-way conceptual components exist, it extracts a compound word: "meaning processing". Likewise, in case of "voice dialogue processing" it extracts keywords as follows if three conceptual components ("voice", "dialogue", and "processing") exist in text at the same time:

- PR (meaning processing) = Concept (meaning) + Concept (processing)
- PR (spoken dialogue process) = Concept (spoken) + Concept (dialogue) + Concept (processing)

Compound words produced under the production rule are called candidate keywords. This production rule is used to restrict the generation of keyword candidates unrelated to the context of the document, as shown in Fig. 1.

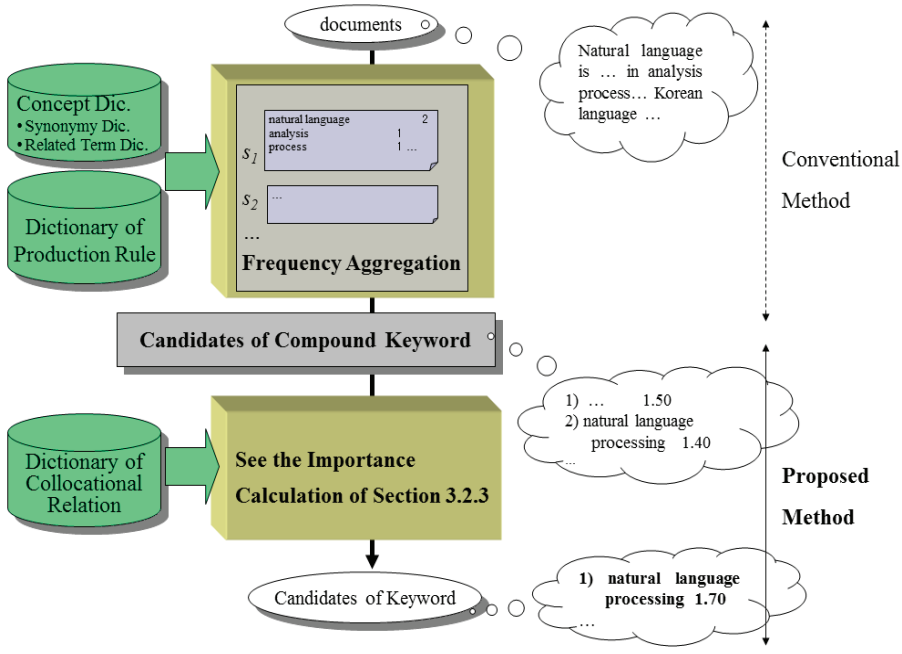

Fig. 1. System architecture.

3.2 Importance of Keyword Candidates

A new method for importance calculation is needed to enhance the extraction's precision level. Sentential distance, *sd*, and conceptual distance, *cd*, are also required to measure importance.

3.2.1 Conceptual distance

A distance between sentences that include each component can be used as a distance between conceptual terms and serve as an indicator of importance. Therefore, the conceptual distance is obtained using the co-occurrence of concept words. However, the distance alone is not enough to understand the relevance between conceptual terms. Therefore, a distance between concepts should be used to focus on the concurrence1 of conceptual terms. For example, let us take a look at PR(XZ) = Concept(X) + Concept(Z).

Distance between sentences, *sd*, is shown as the arrow in Fig. 2. Variables v, x, y, and z are a top-layer language appearing in the document and displayed in lowercase English as conceptual components. Concept words are marked by capitalized alphabets V, X, Y, and Z. The conceptual distance between X and Z is simply 5. *sd* from sentence i to j is defined by $s(j) - s(i) + 1$, (where $j \geq i$).

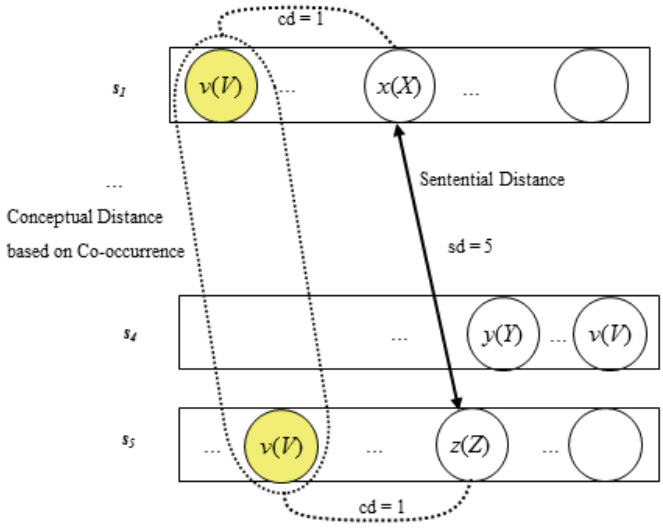

Fig. 2. The relation of sentential distance and conceptual distance.

The conceptual terms are expressed in uppercase English *V*, *X*, *Y*, and *Z*; the conceptual distance between *X* and *Z* is simply the distance between sentences, which is 5 (=5–1+1); and *sd* between sentences *i* to *j* is defined as $s(j) - s(i) + 1$ ($j \geq i$). However, if you focus on the co-occurring concept term *V*, which appears commonly in *X* and *Z*, *V* is in between *X* and *Z*, and the conceptual distance *cd* becomes 1. *cd* is calculated using the following equation.

$$cd = \frac{(cd_1 + cd_2 + \cdots + cd_n)}{cc} \tag{1}$$

where, *cc* is the number of common conceptual word *X*, *cd* of *XV* is 1, and *cd* of *VZ* is 1.

Therefore, *cd* becomes 2(($= \frac{1+1}{1}$, *XV* and *VZ*). This is a calculation of the conceptual distance considering meaningful relevance.

3.2.2 Number of co-occurrences (word sharing)

Conceptual terms or conceptual components that represent the theme often appear in a document. Such concept words are related by the co-occurrence with other words. These terms are also in the same context as the other words in many cases. In this regard, conceptual terms having many words sharing the same context are crucial or most likely to reflect the theme of the entire document. Therefore, it is essential to use a concept word that has a large number of co-occurrences.

Fig. 3 demonstrates the co-occurrences in the document where concept words with co-occurrences are connected by a dashed line. If, in some document, the *i*-th complex word *w* has the number of co-occurrences *N*(*wi*), the number of co-occurrences of *i*-th *w* is shared with other word groups. Concept word *V* and *N*(*V*) become 3(=1+1+1) as in Fig. 4, which shows the number of words involved in the concurrence relationships of *V*. Also, because *V* is the only conceptual term in a concurrent relationship with *X*, *Y*, and *Z*, *N*(*X*), *N*(*Y*), and *N*(*Z*) all become 1. In conclusion, $N(V) > N(X)$, $N(Y)$, or $N(Z)$ and *V* has higher importance than *X*, *Y*, and *Z*.

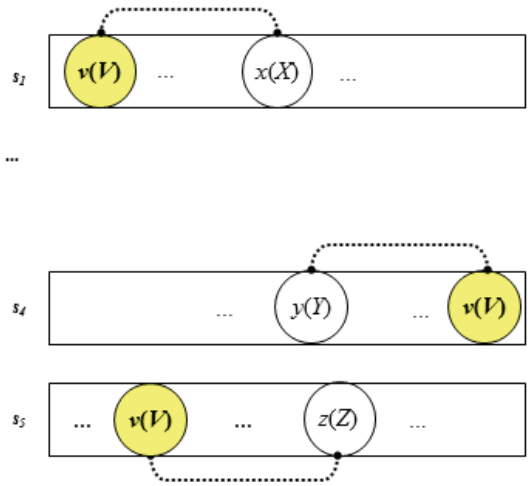

Fig. 3. Concept word with co-occurrence.

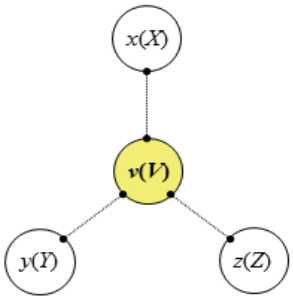

Fig. 4. Number of co-occurrences of concept word *V*.

3.2.3 Calculation of importance

Calculation of the importance of keyword candidates that consider the conceptual distance (*cd*) and the number of concurrences of a conceptual term *wi*, or *N*(*wi*), can be expressed as in Eq. (2). The smaller *cd*, the larger the number of concurrences and the higher frequency of the higher importance, *I*. Furthermore, the smaller *cd*, the higher frequency of synonyms (S) and related terms (R), of concept elements and the higher the importance. The following Eq. 2 expresses how to calculate the importance of candidate keywords considering the distance between conceptual terms *cd* and the number of co-occurrences of a concept word, *N*(*wi*); this is expressed in Eq. (2).

$$I = \left[\frac{1}{n \times cd}\right] \times \sum_{i=1}^{n} \left[\left\langle \frac{\{(S(w_i) \times \alpha) + (R(w_i) \times \beta)\}}{(S_T \times \alpha) + (R_T \times \beta)} \right\rangle \times N(w_i)\right] \quad (2)$$

Here, *n* is the number of conceptual terms that make up the keyword candidates, *S*(*wi*) is the frequency of synonyms of *wi*, *R*(*wi*) is a frequency of how ofter *wi*, related terms appear *ST* is the overall frequency of the synonyms, *RT* is the overall frequency of the related terms, α and β are represent weighted values of synonyms and related terms, respectively (where, $\alpha > \beta$).

In Table 2, the underlined words in the sample documents are words of interest and concept words that appear in the document. The distribution of these words is as shown in Table 3. Fig. 5 shows the co-occurrence relation between the concept element and concept keyword. Next, the production rule calculates the keyword importance by synonyms and similar meaning word candidates of extracted concept keywords. For example, the following "natural language processing," in example (a) has n of 2 ("language", "natural language"), S(natural language) of 6, S(processing) of 5, and, therefore, ST becomes 17(=2+6+5+2+1+1). N(natural language) is 5, N(processing) is 4, and cd is the minimal distance 1. Therefore, the importance is calculated as follows (where, the weighting of synonyms and related terms are respectively $\alpha = 1$ and $\beta = 0.5$; also, the synonyms of this concept word are assumed to not appear in the document).

- PR (natural language processing) = Concept (natural language) + Concept (processing)
- PR (natural language interpretation) = Concept (natural language) + Concept (interpretation)
- PR (Korean language interpretation) = Concept (Korean language) + Concept (interpretation)

Table 2. Sample document ① (<KYWD>>Natural Language Processing // Processing Undefined Words // GLR Rule)

Sentence no.	Example sentence
s_1	It is one of the inspection methods of processing undefined words using an MSLR parser.
s_2	There is a processing problem in natural language processing based on the CFG model utilizing a dictionary, and it is that words not found in the dictionary and undefined words are not attributed to part of speech through terminal symbols.
s_3	On the other hand, there is the LR (GLR) method, which is a more effective method of natural language processing.
s_4	A phoneme of undefined word and breakdown writing… studies on words using the GLR rule… studies on processing undefined Korean words are not practical.
s_5	This study investigates the processing of undefined Korean words according to the GLR rule.
s_6	The experiment was performed using the MSLR parser expanded from the GLR and ETRI electronic dictionary … by restraining the LR table.

Table 3. Conceptual words and conceptual elements extracted from sample document ① in Table 2

Sentence	Conceptual words					
	Analysis	**Natural language**	**Processing**	**Korean**	**Method**	**Model**
s_1	-	Word	Processing	-	-	-
s_2	Analysis	Natural language, word	Processing, processing	-	-	Model
s_3	Analysis	Natural language	-	-	Method	-
s_4	-	Word	Processing	Korean	-	-
s_5	-	Word	Processing	Korean	-	-
s_6	-	-	-	-	-	-
S(wi)	2	6	5	2	1	1

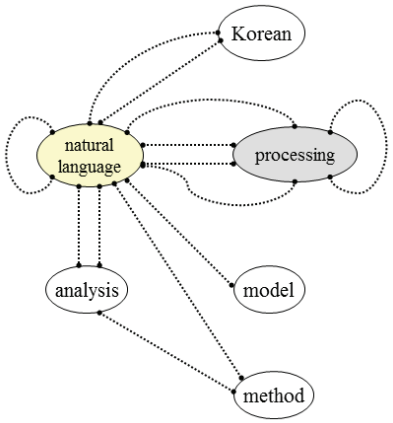

Fig. 5. Co-occurrence relation between the concept element and concept word.

Conceptual words

$$I = \left[\frac{1}{n \times cd}\right] \times \left\{\left\langle \frac{\{(S(natural\ language) \times \alpha) + (R(natural\ language) \times \beta)\}}{(S_T \times \alpha) + (R_T \times \beta)} \right\rangle \times N(natural\ language)\right\} + \left\{\frac{\{(S(processing) \times \alpha) + (R(processing) \times \beta)\}}{(S_T \times \alpha) + (R_T \times \beta)} \times N(natural\ language)\right\}$$
$$= \left[\frac{1}{2 \times 1}\right] \times \left\{\left\langle \frac{\{(6 \times 1) + (0 \times 0)\}}{(17 \times 1) + (0 \times 0)} \right\rangle \times 5\right\} + \left\{\frac{\{(5 \times 1) + (0 \times 0)\}}{(17 \times \alpha) + (0 \times 0)} \times 4\right\} = \frac{1}{2} \times \frac{50}{17} \cong 1.47$$

"Natural language interpretation" of (b) is

$$I = \left[\frac{1}{2 \times 1}\right] \times \left\{\left\langle \frac{6}{17} \right\rangle \times 5\right\} + \left\{\frac{2}{17} \times 4\right\} = \frac{23}{17} \cong 1.35,$$

and "Korean language interpretation" of (c) is

$$I = = \left[\frac{1}{2 \times 1}\right] \times \left\{\left\langle \frac{2}{17} \right\rangle \times 2\right\} + \left\{\frac{5}{17} \times 4\right\} = \frac{12}{17} \cong 0.70$$

If we calculate the importance of the rest of the keyword candidates, then the keyword candidate with the greatest number of co-occurrences is "natural language processing." This keyword candidate also has the greatest importance. This shows that the author-provided keyword "natural language processing" at the bottom of Table 2 (marked by <KYWD>>) is the most appropriate keyword that captures the main topic of the document. It can also be conveyed from calculation results that the keyword "natural language interpretation" with importance greater than 1 is also an appropriate keyword candidate. Fig. 5 shows a graph for Sample ① about the co-occurrence relation between the concept element and concept word.

Next is an example of keyword generation using a group of related terms, as in Table 4 of sample document ②. The main topic of the text is termed as "understanding acoustic environment". The appearance of the synonym "sound" and related term "voice" produces the concept word "acoustic", and the appearance of the synonym "environment" generates the concept word "environment". Lastly, no synonyms for the word "understanding" appeared, but a concept element "recognition" in related terms appears three times to generate the word "understanding". Conclusively, the compound keyword "understanding acoustic environment" (different from the author provided keyword <KYWD>>) is formed from "understanding + acoustic + environment".

Table 4. Sample document ② (<KYWD>>Speech Recognition // Acoustic Environment Recognition)

Sentence no.	Example sentence
s_1	First, the sound stream is segregated to evaluate the speech recognition system.
s_2	This study discusses the problem of using sound stream segregation as a preprocessor of the speech recognition system and prepares pre-experiments.
s_3	… segregation of sound stream transforms the spectrum of the input sound.
s_4	There are wave structure extraction, the transfer function of the initial sound, and the grouping for the transformation.
s_5	Investigate spectrum transformation for the scattered and uniform codebook-type HMM-LR, and soundwave structure extraction barely influences speech recognition…

4. Extraction and Evaluation

4.1 System Configuration

The schematic of the keyword generation system based on the production rule is illustrated in Fig. 1. This system consists of two core modules of keyword generation and importance calculation. The keyword generation part selects only nouns from words found in the dictionary, and finds synonyms and related terms for concept elements. It then uses the production rule to create compound keywords as keyword candidates. The importance attribution part applies synonym and related term weighting, and uses co-occurrence information in the importance equation in Section 3.2.3 to extract keywords in the order of greatest importance.

4.2 Evaluation of Author-Provided Keywords

The experiment of this paper is based on 5,400 sets of titles and summaries of master's and PhD journal articles (about 28.5 kB) offered by KTSET (Korean Test Set) and Jeonju University's academic information center [15]. Table 5 shows keywords extracted from the system and those that were unsuccessfully detected. The experiment considered author-provided keywords but did not appear in the titles or abstracts as answer keywords and conducted the extraction.

Of 2,010 (average number of 4 author keywords in 1 set made up of a title and summary) author-provided keywords, there were 1,540 answer keywords. About 70% of the total keywords were compound keywords, and it had an average of two composite morphemes. This shows that people often use compound keywords to capture the topic of a document. There were 114 keywords extracted using the production rule and 500 keywords generated using the dictionary (case #6 in Table 1), totaling 2,190. A total of 3,350 keywords were not extracted, which means that 17.5% of the papers did not have enough trails to generate keywords, and 47.5% lacked the concept elements (synonyms, related terms) in the dictionary. The production rule was not available at all for the 35%.

Recall and precision of author keywords were evaluated using Eqs. (3) and (4). The recall was the number of extracted answer keywords over the total number of answer keywords, and the precision was the number of extracted answer keywords over the total number of extracted keywords. The recall was 35%, while the precision was 11%. The noticeably low precision is likely caused by the fact that the

number of answer keywords is only 1, whereas the number of author-provided keywords is 4. Therefore, keyword validation is necessary.

$$Recall = \frac{No.\ of\ Extracted\ Answer\ Keywords}{Total\ Answer\ Keywords}. \tag{3}$$

$$Precision = \frac{No.\ of\ Extracted\ Answer\ Keywords}{No.\ of\ Extracted\ Keywords}. \tag{4}$$

Table 5. Comparison of both keywords extracted by system and author-provided keywords (<KYWD>>)

Kinds of keywords	Number of answer keywords
Keywords extracted	1,900
By the production rules	1,400
By the dictionary (case #6 in Table 1)	500
Keywords not extracted by system	3,500
No. of total keywords	5,400

4.3 Keyword Validation

Five experimenters evaluated the validation of the rule-based extraction of keywords using the rating in Table 6. A keyword is determined to be valid if more than four evaluators scored it with B or above, and then it is assumed as an answer keyword. The accuracy rate was defined as the number of answer keywords over the number of extracted keyword candidates as in Eq. (5). The test results were compared to that of Nagata and Kimoto [6].

$$Accuracy\ Rate(\%) = \frac{No.\ of\ Answer\ Keywords}{No.\ of\ Extracted\ Keyword\ Candidates} \times 100 \tag{5}$$

First, let us look at Nagata's experiment in Fig. 6 (labeled as the previous method of keyword extraction). It first defines the index dictionary that states the key concepts and keyword relationships and uses concepts that appear in the document to generate keywords. It also uses Eq. (6) to calculate the importance. The major variables of this equation are the number of concept words that form keyword candidates, conceptual distance based on co-occurrence, and frequency.

$$Importance = \frac{1}{No.\ of\ Concept\ Words\ from\ Keyword\ Candidates} \times \frac{Frequency}{Sentential\ Distance} \tag{6}$$

As seen in Fig. 6, precision improved up to 96%, which is a result of comparing the top seven candidate keywords among extracted ones with that of Nagata and Kimoto [6]. The experiment selected 30 words based on field terms and thematic words. A dictionary [8] classified by Kadogawa was used to refer to conceptual components after being translated into Korean. For some terminologies related to natural language and auditory language that were not listed on the dictionary, people manually selected words from the classification chart by matching the closest meaning as best as possible and used synonyms and the related terms as the conceptual components [15-17]. For the weight value of synonyms and related terms, multiple preliminary experiments were conducted, and the optimal value was obtained: α = 1 and

β = 0.5, which were the optimum number determined from pre-experiments and a heuristic algorithm [18].

Table 6. Validity of keyword candidates

Rating	Validity
A	Valid as a keyword
B	Marginally valid as a keyword
C	Slightly invalid as a keyword
D	Invalid as a keyword

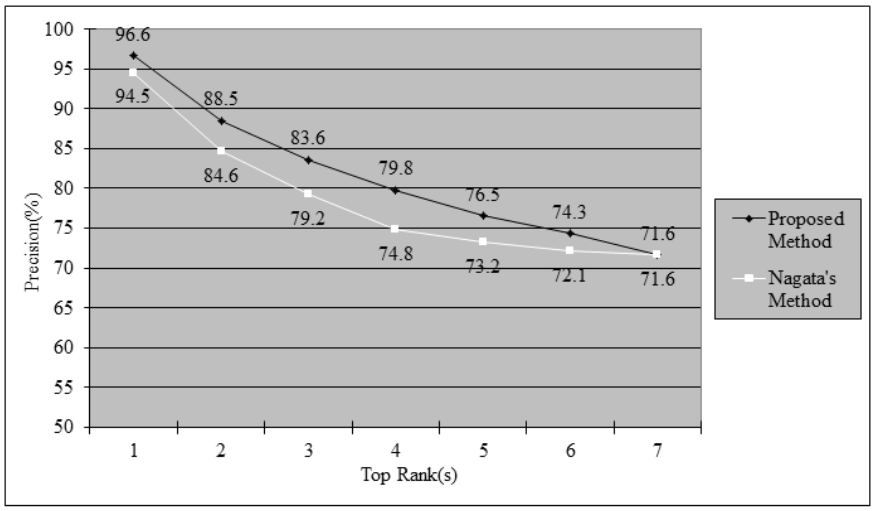

Fig. 6. Comparison of proposed method and Nagata's method.

5. Conclusions

The present system suggests proper keywords to assist in the quick determination of whether they will be read or not by informing them of keywords having the highest importance. This is a new introduction of the concept-based compound keyword extraction focused on extracting inference words [14] or thematic terms to help readers easily guess the document category [19,20]. This study defines a concept-based keyword production rule to improve the keyword precision and introduces a method to calculate the keyword importance by using the frequency of word appearance, concurrence, and distance between concepts. One of the greatest points of this method is that keywords that did not appear in a document can be extracted as well as author keywords. To improve the accuracy rate of extracted keywords, it defines a concept-based production rule and proposes a calculation method of keyword importance using frequency count, co-occurrence, and conceptual distance. The biggest advantage of the developed method is that it can extract not only the author-provided keywords but also keywords that do not appear in the text body or database [21] of the document.

This study's limitations include the method's inability to extract field terminologies when they are not listed in the dictionary of concept elements. Thus, for future studies, the dictionary of concept elements

can be established to increase the accuracy rate, and it can further be applied to the information retrieval system currently in development. Moreover, several studies [22-24] can be beneficial for future research [10,21-23,25].

References

[1] H. Kimoto, "Automatic indexing and evaluation of keywords for Japanese newspapers," *IEICE Transactions on Information and Systems, Pt.1 (Japanese Edition)*, vol. 74, no. 8, pp. 556-566, 1991.

[2] H. Kimoto, "Automatic indexing of an integrated large scale text database and its evaluation," *IPSJ SIG Technical Reports*, vol. 92, no. 71 (DBS-90), pp. 73-81, 1992.

[3] H. Suzuki, S. Masuyama, and S. Naito, "Examination of keyword extraction using thesaurus in Japanese text," *IPSJ SIG Technical Reports*, vol. 93, no. 101 (NL-98), pp. 73-80, 1993.

[4] K. Uchiyama and M. Nakamura, "Development of an automatic keyword-extracting system on the basis of content analysis and an application system," *IPSJ Research Report Database System*, vol. 1991, no. 65 (DBS-084), pp. 151-160, 1991.

[5] M. Okumura and H. Nanba, "Automated text summarization: a survey," *Journal of Natural Language Processing of Japan*, vol. 6, no. 6, pp. 1-26, 1999.

[6] M. Nagata and H. Kimoto, "A newspaper keyword generation method based on key concept extraction," *Proceedings of the 37th National Convention Information Processing Society of Japan*, Tokyo, Japan, 1988, pp. 1030-1031.

[7] N. Kando, K. Kuriyama, T. Nozue, and K. Oyama, "NTCIR-1 (NACSIS Test Collection for Information Retrieval systems-1): Its Policy and Practice," *IPSJ SIG Technical Reports*, vol. 99, no. 20, pp. 33-40, 1999.

[8] S. Ono and M. Hamanishi, *Kadokawa Ruigo Shin Jiten*. Tokyo, Japan: Kadokawa Shoten, 1981.

[9] M. Hara, H. Nakajima, and T. Kitani, "Keyword extraction using a text format and word importance in a specific field," *IPSJ Journal*, vol. 38, no. 2, pp. 299-309, 1997.

[10] M. Morohashi, "Automatic indexing survey," *IPSJ Magazine*, vol. 25, no. 9, pp. 918-925, 1984.

[11] S. Ito, H. Niwa, K. Kayashima, S. Maruno, and Y. Shimeki, "Parametric keyword extraction algorithm and adaptation method," *IEICE Technical Report (Natural Language Understanding and Models of Communication)*, vol. NLC93-53, pp. 41-46, 1993.

[12] T. Tokunaga, *Information Retrieval and Natural Language Processing*. Tokyo, Japan: University of Tokyo Press, 1999.

[13] Y. Ogawa, M. Mochinushi, and A. Bessho, "A compound keyword assignment method for Japanese texts," *IPSJ SIG Notes*, vol. 93-NL-97-15, no. 9, pp. 103-110, 1993.

[14] S. S. Lee, M. Shishibori, T. Sumitomo, and J. I. Aoe, "Extraction of field-coherent passages," *Information Processing & Management*, vol. 38, no. 2, pp. 173-207, 2002.

[15] Y. H. Chen, E. J. L. Lu, and M. F. Tsai, "Finding keywords in blogs: efficient keyword extraction in blog mining via user behaviors," *Expert Systems with Applications*, vol. 41, no. 2, pp. 663-670, 2014.

[16] J. A. L. Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Towards a Mixed Approach to Extract biomedical terms from text corpus," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, vol. 4, no. 1, pp. 1-15, 2014.

[17] M. Dostal and K. Jezek, "Automatic keyphrase extraction based on NLP and statistical method," in *Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Objects*, Pisek, Czech Republic, 2011, pp. 140-145.

[18] O. Mirzaei and M. R. Akbarzadeh-T, "A novel learning algorithm based on a multi-agent structure for solving multi-mode resource-constrained project scheduling problem," *Journal of Convergence*, vol. 4, no. 1, pp. 47-52, 2013.

[19] A. Buschettu, D. Sanna, G. Concas, and F. E. Pani, "A platform based on Kanban to build taxonomies and folksonomies for DMS and CSS," *Journal of Convergence*, vol. 6, no. 1, pp. 1-8, 2015.
[20] R. Al-Hashemi, "Text Summarization Extraction System (TSES) using extracted keywords," *International Arab Journal of e-Technology*, vol. 1, no. 4, pp. 164-168, 2010.
[21] Y. L. Choi, W. S. Jeon, and S. H. Yoon, "Improving database system performance by applying NoSQL," *Journal of Information Processing Systems*, vol. 10, no. 3, pp. 355-364, 2014.
[22] R. Benlamri and X. Zhang, "Context-aware recommender for mobile learners," *Human-centric Computing and Information Sciences*, vol. 4, article no. 12, 2014.
[23] H. Im, J. Kang, and J. H. Park, "Certificateless based public key infrastructure using a DNSSEC," *Journal of Convergence*, vol. 6, no. 3, pp. 26-33, 2015.
[24] T. Kwon, J. Lee, H. Choi, O. Yi, and S. Ju, "Efficiency of LEA compared with AES," *Journal of Convergence*, vol. 6, no. 3, pp. 16-25, 2015.
[25] N. Katoh and N. Uratani, "A new approach to acquiring linguistic knowledge for locally summarizing japanese news sentences," *Journal of Natural Language Processing*, vol. 6, no. 7, pp. 73-92, 1999.

Samuel Sangkon Lee **https://orcid.org/0000-0001-9965-8387**

He received his B.S. and M.S. degrees from the Department of Computer Science of Chonbuk National University, South Korea, in 1996 and 1998, respectively. He then received his Ph.D. from the Department of Information Science and Intelligent Systems of Tokushima University, Japan, in 2001. Since 2002, he has been a professor in the Department of Computer Science and Engineering at Jeonju University, South Korea. His research interests include information retrieval, keyword extraction, document classification, and natural language processing.