ISSN 2007-9737

Relation between Titles and Keywords in Japanese Academic Papers using Quantitative Analysis and Machine Learning

Masaki Murata, Natsumi Morimoto

Tottori University, Faculty of Engineering, Tottori, Japan

{murata, s112057}@ike.tottori-u.ac.jp

Abstract. In this study, we analyzed keywords from different academic papers using data from more than 300 papers. Using the concept of quantitative surveys and machine learning, we conducted various analyses on the keywords in different papers. The findings obtained from these surveys and analyses are assumed to lend themselves to the automatic assignment of keywords for papers. In this study, the number of keywords included in a paper is quantitatively expressed using the covering rate and density of keywords. The results confirm that paper titles are likely to include keywords. The performed keyword analyses predict words that can be used as keywords via machine learning. The proposed method has an accuracy range 0.6–0.8. In addition, by analyzing the features used in machine learning, we can obtain the characteristics of the words that are mentioned as keywords in papers.

Keywords. Thesis, title, keyword, machine learning, feature analysis

1 Introduction

In recent years, along with the enormous increase in the number of electronic documents, the importance of developing a method to facilitate the retrieval of information from large amounts of documents has increased. Many journals require authors to provide keywords for papers.

Nagao [5] proposed a search method that uses the table-of-contents information of books and documents. Nagao mentioned that for academic papers, the number of keywords attached by the author does not sufficiently reflect the content of the paper.

He stated that because the titles of papers and the titles of chapters and sections have an appropriate number of technical terms suitable for expressing the content of a paper, it is reasonable that these titles be used as paper keywords.

Nagao used 16 papers for their analysis. Conversely, in this study, we expand the study scale by using a large amount of data, i.e., 300 papers or more, and primarily analyze the keywords of the papers. We investigate whether the titles in a paper can be used as keywords of the paper and examine the characteristics of the keywords.

The obtained findings will likely lead to the automatic generation of keywords for papers and an improvement in the keyword generation performance.

The highlights of this study can be summarized as follows:

— In this study, we conducted a survey and analysis of keywords in academic papers using a large amount of data. The covering rate of how much the title of a paper covers the keywords given by the author and the density of the keywords the author provided in the title of the paper are investigated. The fact that paper titles tend to include keywords was confirmed.

— We conducted experiments predicting the words that could be keywords of a paper using the titles of papers, its chapters, and its sections via supervised machine learning; our proposed methods obtained

an accuracy rate of 0.6 for the prediction of keywords. In other experiments wherein human beings determined the correct answers in an evaluation, our proposed methods obtained an accuracy rate of 0.8 for the prediction of keywords.

— By analyzing the features used for supervised machine learning, we easily obtained the characteristics of words including “analysis” and found that words in titles tend to be keywords and that words such as “morphological analysis” and “syntactic analysis” tend to be keywords.

2 Previous Studies

Nagao [5] proposed a method to select keywords from the table-of-contents information and to search books and documents using the structures of the chapters and sections in the table of contents when creating a retrieval system for books and documents.

He manually investigated 16 academic papers. As a result, for academic papers, he concluded that the number of keywords attached by the authors is inadequate and does not fully reflect the contents of a paper. He stated that because the titles of papers and the titles of chapters and sections include a sufficient number of technical terms to express a paper, it is reasonable for these titles to be used as the keywords of a paper.

Kurohashi et al. [2] examined the strength of the relations among terms in the text and the table of contents using the hierarchically structured information in the table of contents and made it possible to conduct a document search based on not only the conventional AND and OR connections but also the degree of keyword connections. In an evaluation experiment to investigate the effectiveness of these methods, they assessed approximately 17,000 volumes of table-of-contents information. As a result, they quantitatively proved that their proposed method was effective in book searches by evaluating 80 search results by subject.

Nagao and Kurohashi et al. showed that the table-of-contents information can effectively serve as keywords in book searches. In this study, we also investigated whether the table-of-contents information represents the characteristics of academic papers but used a method different from those of Nagao and Kurohashi et al.

Uchiyama et al. [7] proposed a method to extract keywords from abstracts via statistical methods, using keywords given by the authors of specialty papers as learning data for the keywords. Uchiyama et al.’s proposed method obtained a recall of 0.8 and a precision of 0.43, indicating the effectiveness of their method. Uchiyama et al. extracted keywords given by authors from summaries using statistics. Conversely, in this study, we primarily aim to predict keywords from titles using supervised machine learning. In addition, Uchiyama et al. used the words immediately before and after the target word to obtain statistics for the learning data. Conversely, in this study, learning data are created by using not only words immediately before and after a target word but also the information contained in additional words further from the target.

In addition to the above, many studies concerning keyword extraction have been conducted. Research concerning keywords such as titles and abstracts have also been conducted [1]. In this study, machine learning is used to determine keywords using the frequency of a word, the importance of the sentence in which the word exists, and the length of the word. In our study, the words before and after a word as well as a thesaurus are used as information for the learning process and a machine learning method is used to predict the keywords.

In addition, there are studies [8] that have extracted keywords from documents using multiple machine learning techniques. In our study, machine learning is used as well; however, the keywords are estimated not only from the entire document but also from the table-of-contents information using paper titles and their chapter and section titles, and an analysis of the features is performed. In our study, we focused on the importance of titles.

ISSN 2007-9737

3 How to Proceed with this Research

This study was conducted in the following order:

— Quantitative survey of keywords

We quantitatively described how many keywords were included in a paper using the covering ratio and the density of keywords and conducted surveys on the keywords.

— Analysis of keywords based on machine learning

By predicting which words could become keywords using supervised machine learning techniques such as the maximum entropy method and support vector machine *SVM* and analyzing the features that were useful for the machine-learning-based prediction, it was possible to evaluate the characteristics of words that are likely or unlikely to be keywords.

4 Quantitative Survey of Keywords

As mentioned in Section 2, Nagao [5] determined that the keywords given by an author do not adequately reflect the content of a paper. However, the keywords given by an author (which we call author keywords) are important words in the paper and can be used as a certain index for the keywords.

We compared the author keywords in a paper with the following elements of the paper and investigated how many keywords were included in the paper or in each part of the paper:

— Paper title,

— Titles of chapters and sections,

— Titles of papers, chapters, and sections,

— Abstract,

— Titles of papers, chapters, and sections and abstract,

— Entire paper.

4.1 Used Data

In this study, we used 343 papers in the Japanese Journal of Natural Language Processing (over a period of 16 years) as experimental data.

A paper in this dataset contains a title, an abstract, keywords, chapter/section titles, and the text. Each paper has approximately five keywords.

4.2 Survey Method

The covering ratio of author keywords is defined as the number of phrases in an element matching the author keywords divided by the number of author keywords. This ratio was examined for each article, and the average value of the ratio was obtained.

The covering ratio of words in the author keywords is defined as the number of words in an element matching the author keywords divided by the total number of words in the author keywords. This ratio was examined for each article, and the average value of the ratio was obtained. To partition an author keyword into words, we used *ChaSen*[1].

The density of words in the author keywords is defined as the number of words in an element matching the words in the author keywords divided by the number of words in the element. This ratio was examined for each paper, and the average value of the ratio was obtained.

An example of calculating the covering ratio of the author keywords, the covering ratio of the words in the author keywords, and the density of words is shown in Fig. 1.

4.3 Investigation Result

Table 1 shows the results for the covering ratio of author keywords. Table 2 shows the results for the covering ratio of the words in author keywords. Table 3 shows the results for the density of words in author keywords.

[1] http://chasen-leagacy.sourceforge.jp/

1. Author keywords
   Machine learning, case analysis, and support vector machine
2. Title
   Case analysis using machine learning
3. Covering ratio of author keywords
   2 (the number of matches) / 3 (the number of author keywords)
4. Covering ratio of the words in author keywords
   4 (the number of matches) / 7 (the number of author keywords)
5. Density of words
   4 (the number of matches) / 5 (the number of words in the title)

Fig. 1. Examples of the calculations of covering ratios and density

Table 1. Results for the covering ratio of author keywords

Element	Results
Title	0.36
Chapter/Section titles	0.48
Title and chapter/section titles	0.58
Abstract	0.60
Title, chapter/section titles, and abstract	0.71
Entire paper	0.86

4.4 Discussion

First, we numerically compared the results of the covering ratio of author keywords with those of the words in the author keywords. Looking at Tables 1 and 2, the covering ratio of the words in author keywords is approximately 0.2 higher than the covering ratio of author keywords. In the case of the author keywords as is (Table 1), the covering ratio of the titles is approximately 0.4, the covering ratio of the titles and chapter/section titles is approximately 0.5, and the covering ratio of entire papers is approximately 0.9. However, in the case of splitting author keywords into words (Table 2), the covering ratio of the titles is approximately 0.5, the covering ratio of the titles and chapter/section titles is approximately 0.7, and the covering ratio of entire papers is approximately 1.0.

Next, when analyzing the results of the covering ratio of the words in author keywords, words that are not keywords themselves were found to be included in the keywords. For example, we considered the keyword *LR koubun kaiseki* "LR syntax analysis." Even though this keyword may not appear in an entire given paper, the individual word groupings *LR* "LR" and *koubun kaiseki* "syntax analysis" do appear in such a paper.

Table 2. Results for the covering ratio of the words in author keywords

Element	Results
Title	0.53
Chapter/Section titles	0.67
Title and chapter/section titles	0.76
Abstract	0.82
Title, chapter/section titles, and abstract	0.89
Entire paper	0.98

Table 3. Results for the density of the words in author keywords

Elements	Results
Title	0.41
Chapter/Section titles	0.22
Title and chapter/section titles	0.25
Abstract	0.13
Title, chapter/section titles, and abstract	0.16
Entire paper	0.084

Finally, we analyzed the results of the density of words in the author keywords. From Table 3, the density of the entire paper is less than 0.1; however, the density of the titles exceeds 0.4, which is relatively high. The second highest density is that of the titles and chapter/section titles. If words are randomly extracted from an entire paper, the accuracy rate of extracting the correct words from the author keywords is approximately 0.1. However, if words are randomly extracted from a paper title, the accuracy rate of extracting correct words from the author keywords is approximately 0.4. Words from a paper title are therefore more likely to be words in the author keywords than words from the entire paper.

From the above, it can be concluded that words in paper titles or chapter/section titles tend to be keywords.

ISSN 2007-9737

5 Analysis of Keywords based on Machine Learning

Section 4 demonstrated that words that could be keywords were contained in the paper titles. However, unnecessary words that are not keywords are included in titles. Machine learning was performed to judge whether the detected words could be keywords of a given paper. We analyzed the differences between the words that could be keywords and the words that could not.

5.1 Our Proposed Method

In this study, we used supervised machine learning to predict which words could be used as keywords of articles. We separate the elements of the paper (i.e., the paper title, chapter title, section title, abstract, and whole paper) into words via *ChaSen* and input them one by one into the machine. The machine judges whether the input word can be a keyword and outputs the result.

In this study, SVM and the maximum entropy method were used for machine learning in our experiments. SVM supports high-performance supervised machine learning and has been used in many studies. The maximum entropy method automatically finds features with high weights, with features easier to analyze than those of SVM. Thus, it was also used for our experiments.

The features used in the machine learning are shown in Table 4. The features that use *Bunrui Goi Hyou* (a Japanese thesaurus) [6] are defined by consulting Murata's method [3] and can utilize semantic categories (e.g., human beings and animals).

Table 4. Features used in the machine learning

ID	Feature
1	Element where the target word appears (title, chapter title, section title, abstract, and text)
2	The target word
3	One word just before the target word
4	Two words just before the target word
5	Three words just before the target word
6	One word just after the target word
7	Two words just after target word
8	Three words just after target word
9	Chapter number and section number
10	The three first digits in *Bunrui Goi Hyou* (a Japanese thesaurus)
11	The five first digits in *Bunrui Goi Hyou*

5.2 Baseline Methods

We explain the baseline methods that were used in the experimental comparison in this section.

A Method using all Words as Keywords

In this study, a morphological analysis was performed using *ChaSen* and the elements of the papers (i.e., paper title, chapter/section title, abstract, and entire article) were divided into words. The method "a method using all words as keywords" considers all words as keywords and was used as one of the comparative methods. This method corresponds to a case where there is no information (features), and it was used to evaluate whether the machine learning technique works well.

A Method using all nouns as Keywords

The elements of the papers (i.e., paper title, chapter/section title, abstract, and entire article) were divided into words. The method "a method using all nouns as keywords" considers all nouns as keywords and was also used as a comparative method.

Gensen Web

We extracted the keywords of a paper using an automatic technical term extraction service called "Gensen Web."[2]

In Gensen Web, when we enter a text, the technical terms (keywords) are displayed in descending order of importance. We input 343 papers and obtained the keywords for each paper. We picked out the words appearing in the elements of the papers from the keywords obtained by Gensen Web.

We sorted the words appearing in the elements obtained by Gensen Web in descending order of importance. We selected the top 7 words for the paper title; the top 8 words for the chapter/section title; the top 9 words for the paper title and chapter/section title; the top 12 words for the

[2] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html

Table 5. Number of data items in the training data

Element	Number of papers	Number of target words	Number of words in author keywords	Number of words not in author keywords
Title	171	1628	814	814
Paper/ chapter/ section title	171	2540	1270	1270
Paper/ chapter/ section title and abstract	172	3080	1540	1540

Table 6. Number of data items in the test data

Element	Number of papers	Number of target words	Number of words in author keywords	Number of words not in author keywords
Title	172	2063	775	1288
Paper/ chapter/ section title	172	10577	1183	9394
Paper/ chapter/ section title and abstract	172	23967	1438	22529

abstract; the top 13 words for the paper title, chapter/section title, and abstract; and the top 11 words for the entire paper. We used the selected words as the words that Gensen Web considered as words in keywords.

For the number of words we selected, we used the number of words for which we obtained the highest F-measure in the training data.

The TF–IDF method

In the term frequency–inverse document frequency (TF–IDF) method, we found the TF–IDF value of a word appearing in a paper and extracted words with high TF–IDF values as keywords.

The formula for the TF–IDF method is shown below:

$$tfidf_{i,j} = tf_{i,j} \times idf_i, \tag{1}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{2}$$

$$idf_i = log \frac{|D|}{|d : d \ni t_i|}, \tag{3}$$

here, $n_{i,j}$ is the number of occurrences of a word i in the document j, $|D|$ is the total number of documents (papers), and $|\{d : d \ni t_i\}|$ is the number of documents containing the word i.

We found the TF–IDF values of the words in the 344 papers using the above equation. We extracted the words appearing in the elements of each paper and sorted them in descending order according to their TF–IDF values. We selected the top 8 words for the paper title; the top 13 words for the chapter/section title; the top 12 words for the paper title and chapter/section title; the top 28 words for the abstract; the top 37 words for the

Table 7. Results for extracting keywords from paper titles

Element	Recall rate	Precision rate	F-measure
SVM	0.76 (591/775)	0.57(591/1042)	0.65
Maximum entropy	0.72(560/ 775)	0.57 (560/977)	0.64
All words as keywords	1.00 (775/775)	0.38(775/2063)	0.55
All nouns as keywords	0.85 (658/775)	0.50(658/1316)	0.63
Gensen Web	0.85 (655/775)	0.57(655/1141)	0.68
TF-IDF method	0.75 (584/775)	0.43(584/1354)	0.55

Table 8. Results for extracting keywords from paper/chapter/section titles

Element	Recall rate	Precision rate	F-measure
SVM	0.77 (905/1183)	0.35 (905/2582)	0.48
Maximum entropy	0.77 (913/1183)	0.32 (913/2821)	0.46
All words as keywords	1.00(1183/1183)	0.11(1183/10577)	0.20
All nouns as keywords	0.72 (850/1183)	0.12 (850/6900)	0.21
Gensen Web	0.57 (674/1183)	0.43 (674/1557)	0.49
TF-IDF method	0.32 (384/1183)	0.18 (384/2076)	0.24

paper title, chapter/section title, and abstract; and the top 41 words for the entire paper. We used the selected words as the words that the TF–IDF method considered to be keywords.

For the number of words we selected, we used the number of words for which we obtained the highest F-measure in the training data.

5.3 Experimental Data

We used the data described in Section 4.1. The training data used in the experiment are shown in Table 5, and the test data are shown in Table 6.

With respect to the training data, words not included in the keywords were randomly extracted so that the number of words was the same as the number of words included in the keywords. This was done for the following reasons. If the number of words included in the keywords and the number of words not included in the keywords are very different, there is a possibility that the machine learning will not work well. Therefore, in this study, we prevented this problem by using the same quantity for the number of words included in the keywords and that of words not included in the keywords.

5.4 Experimental Results

The results obtained using the six methods (our proposed methods (machine learning based on SVM and the maximum entropy method), the method judging all entered words to be keywords, the baseline method, Gensen Web, and the TF–IDF method) are shown in Tables 7–9. The recall rates, precision rates, and F-measures for extracting keywords from the titles are shown in Table 7, the results for the paper/chapter/section titles are shown in Table 8, and the results for the paper/chapter/section titles and abstracts are shown in Table 9. The recall rates indicate the ratio of correct words among words appearing in both the given titles and author keywords, and the precision rates of the correct words among the words that are considered by a given method as words in author keywords. The correct words are the words included in the author keywords.

From Tables 7–9, because the F-measures of the proposed methods are higher than those of the method extracting all words as keywords, we determined that the machine learning method worked well for keyword prediction.

Table 7 indicates that there were smaller differences when comparing the method using all nouns as keywords and the proposed method. However, Tables 7 and 9 indicate that the proposed

Table 9. Results for extracting keywords from paper/chapter/section titles and abstracts

Element	Recall rate	Precision rate	F-measure
SVM	0.74(1059/1438)	0.30 (1059/3498)	0.43
Maximum entropy	0.78(1122/1438)	0.28 (1122/4029)	0.41
All words as keywords	1.00(1438/1438)	0.06(1438/23967)	0.11
All nouns as keywords	0.84(1213/1438)	0.09(1213/13979)	0.16
Gensen Web	0.55 (794/1438)	0.35 (794/2249)	0.43
TF-IDF method	0.42 (611/1438)	0.10 (611/6377)	0.16

Table 10. Experimental results using correct answers based on three human beings (paper titles)

Method	Recall rate	Precision rate	F-measure
SVM	0.83(19/23)	0.80 (19/(19+3×1288/775))	0.81
Maximum entropy	0.83(19/23)	0.80 (19/(19+3×1288/775))	0.81
All words as keywords	1.0 (23/23)	0.45(23/(23+17×1288/775))	0.62
All nouns as keywords	0.96(22/23)	0.69 (22/(22+6×1288/775))	0.80
Gensen Web	0.70(16/23)	0.76 (16/(16+3×1288/775))	0.73
TF-IDF method	0.70(16/23)	0.52 (16/(16+9×1288/755))	0.59

method obtained higher F-measures than the method extracting all nouns as keywords. Because the recall rate of the proposed method is higher than that of the method using all nouns as keywords, it is likely that the proposed method can distinguish between words that are keywords and words that are not keywords even for non-nouns. Comparing the F-measures of Gensen Web and those of our proposed methods, the values of our proposed method were nearly equivalent to or a little lower than those of Gensen Web. Gensen Web is popular and is often used in many studies. Because our proposed methods obtained values similar to those of Gensen Web, we confirmed that our methods also work well.

Moreover, when we compared the results of the TF–IDF method to those of the proposed method, we found that the proposed method can remove unnecessary words with higher efficiency.

5.5 Experiments using Human beings for the Evaluation

We examined the experimental results in the previous section and found that there were words that could be used as keywords other than those provided by the authors. Therefore, we conducted experiments using keywords determined by human beings other than the authors.

We randomly picked five papers from the test data used in Section 5.3. We randomly extracted 20 words in author keywords and 20 words from the titles (the paper/chapter/section titles) of the papers. Three human beings read the five papers and guessed whether each of the 40 words was included in the author keywords.[3] A word that two or more people considered to be a keyword was regarded as a correct answer.

We evaluated the six methods using the correct answers based on the evaluation of the three human beings mentioned above. The results are shown in Tables 10 and 11. We calculated the recall rates using the sampling data and the following equation:

$$\frac{Sys_1}{Sys_1 + Sys_2 \times R}, \tag{4}$$

here, Sys_1 is the number of correct answers among the words that a system considered to be keywords, Sys_2 is the number of incorrect answers

[3]The kappa coefficient in the manual estimation of the three human beings was 0.4. This corresponds to “moderate agreement.”

Table 11. Experimental results using correct answers based on three human beings (paper/chapter/section titles)

Method	Recall rate	Precision rate	F-measure
SVM	0.83(19/23)	0.44 (19/(19+3×9394/1183))	0.58
Maximum entropy	0.83(19/23)	0.37 (19/(19+4×9394/1183))	0.52
All words as keywords	1.0 (23/23)	0.15(23/(23+17×9394/1183))	0.25
All nouns as keywords	0.96(22/23)	0.26 (22/(22+8×9394/1183))	0.41
Gensen Web	0.43(10/23)	0.56 (10/(10+1×9394/1183))	0.49
TF-IDF method	0.30 (7/23)	0.31 (7/(7+2×9394/1183))	0.31

Table 12. Results of feature analysis (paper/chapter/section title)

Useful		Not useful	
Feature	Score	Feature	Score
The input word appears in titles	0.93	The succeeding word is “study”	0.15
kaiseki (analysis)	0.86	The previous word is object case particle	0.14
ka (-ization)	0.85	The succeeding word is “method”	0.11

among the words that a system considered to be keywords, and R is the result of dividing the number of words that are not in author keywords by the number of words that are in author keywords in all the test data.

In the results of Tables 10 and 11, we see that our proposed methods (SVM and maximum entropy) obtained higher F-measures (0.8 for the paper titles) than other methods.

5.6 Feature Analysis

In the maximum entropy method, the weights of features are automatically recognized [4]. That is, it is possible to analyze which feature is useful for determining if a word could be a keyword. A feature analysis of the experimental results obtained by the maximum entropy method in Section 5.4 was performed.

Useful features and non-useful features are shown in Table 12. The results were obtained for the experiments including the paper/chapter/title sections. The score in the table indicates the weighting of a feature. A feature with a high score is more useful for predicting keywords.

From Table 12, we found that when an input word exists in a paper title, it tends to be a keyword.

Further, when the input word is “analysis,” it is often a keyword. The terms “morphological analysis,” “syntax analysis,” “relationship analysis,” etc., were determined to be keywords in the academic field of natural language processing.

We also found that when the word after an input word is “method” or “study,” it does not tend to be a keyword. Accordingly, terms such as “experiment method,” “evaluation method,” “previous study,” etc., are often used for chapters/section titles but are not useful as keywords.

Our study found many other similar keyword characteristics.

6 Conclusions

In this study, we investigated and analyzed keywords in academic papers.

The covering rate of how much the title of a paper covers the keywords given by the author and the density of how many keywords the author gave in the title of the paper were investigated. The fact that the paper titles tended to be keywords was confirmed.

Specifically, paper titles covered approximately 0.5 of the keywords given by the author and the chapter/section titles included more than 0.7. Moreover, the density of the paper titles was 0.4 and it was found that keywords can be obtained

with a correct answer rate of 0.4 even if words are randomly selected from the paper title.

We also predicted words that could be the keywords of a paper using supervised machine learning from the various elements of the paper. As a result of keyword prediction from paper titles and paper/chapter/section titles, it was possible to determine words that could be keywords with accuracy rates of 0.6–0.8.

In addition, by analyzing the features used for supervised machine learning, we were able to obtain the characteristics of words that could be keywords of papers.

Specifically, one characteristic of a word that can be a keyword is that it exists in a paper title. It was found that phrases including "analysis," such as "morpheme analysis" and "syntax analysis" in natural language processing research fields, are likely to be keywords. Conversely, it was found that phrases including "method" and "study," such as "evaluation method" and "previous study," are not likely to be keywords.

References

1. **Bhowmik, R. (2008).** Keyword extraction from abstracts and titles. *Proceedings of IEEE SoutheastCon 2008*, pp. 610–617.
2. **Kurohashi, S., Hagiwara, N., & Nagao, M. (1997).** A system for document retrieval by utilizing the table of contents information. *Information Processing Society of Japan, WGFI*, pp. 27–32. (in Japanese).
3. **Murata, M. & Isahara, H. (2003).** Conversion of Japanese passive/causative sentences into active sentences using machine learning. *Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2003, Mexico City, February 2003 Proceedings*, pp. 115–125.
4. **Murata, M., Uchimoto, K., Utiyama, M., Ma, Q., Nishimura, R., Watanabe, Y., Doi, K., & Torisawa, K. (2010).** Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. *Cognitive Computation*, Vol. 2, No. 4, pp. 272–279.
5. **Nagao, M. (1992).** A method of document retrieval by utilizing the table of contents as keys. *Information Science and Technology Association*, Vol. 42, No. 8, pp. 711–718. (in Japanese).
6. **NLRI (1964).** *Bunrui Goi Hyou*. Shuuei Publishing.
7. **Utiyama, M., Murata, M., & Isahara, H. (2000).** Using author keywords for automatic term recognition. *Terminology*, Vol. 6, No. 2, pp. 313–326.
8. **Zhang, C. (2009).** Combining statistical machine learning models to extract keywords from Chinese documents. *Proceedings of the 5th International Conference on Advanced Data Mining and Applications*, pp. 745–754.

Article received on 17/01/2019; accepted on 06/03/2019.
Corresponding author is Masaki Murata.