IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol.004, LIR308
DOI: https://doi.org/10.52731/lir.v004.308

Context-sensitive Classification for Scientific Keywords in Grant Reports

Michiko Yasukawa [*] , Koichi Yamazaki [†]

Abstract

In the task of institutional research (IR), it is important for each university to identify the latest trends in cutting-edge scientific research and to understand its own strengths. The Grant-in-Aid for Scientific Research (KAKENHI), the largest research grant in Japan, makes publicly available the research outline, progress, and keywords of adopted research projects. These open data can be used to analyze research information in IR tasks. Our study in this paper focuses specifically on keyword analysis in research grant reports. Technical terms that describe scientific projects are important clues in analyzing research information. However, state-of-the-art terminology is not easy to process on computers because word occurrences and usages are often polysemous and unpredictable. To deal with this issue, we propose a method for disambiguating keywords by attaching a prefix to each keyword that takes into account the context in which the keyword appears. Such contextual prefixes are expected to enable useful searches for relevant keywords and automatic classification of keywords. Evaluation experiments on real data confirmed the effectiveness of our proposed method.

Keywords: research project management, faculty development, bibliometrics, text analysis

1 Introduction

Scientific keywords in research grant reports are selected by knowledgeable researchers so that each keyword is important and well represent the research topics concerned. By analyzing such keywords, it is expected to gain useful knowledge for recognizing trends in cutting-edge research, identifying the differences between one university's research and others, and developing strategies for management reform in higher education. The keywords in grant reports should represent the uniqueness and characteristics of the research, and in this regard, a high degree of specificity is required. On the other hand, to indicate the universality of the research and the wide range of its application, comprehensiveness is required to be able to relate it to various concepts and meanings. Therefore, there are extreme discrepancies in the frequency of occurrence of words and usage of words, depending on

[*] Gunma University, Gunma, Japan
[†] Tokyo Denki University, Tokyo, Japan

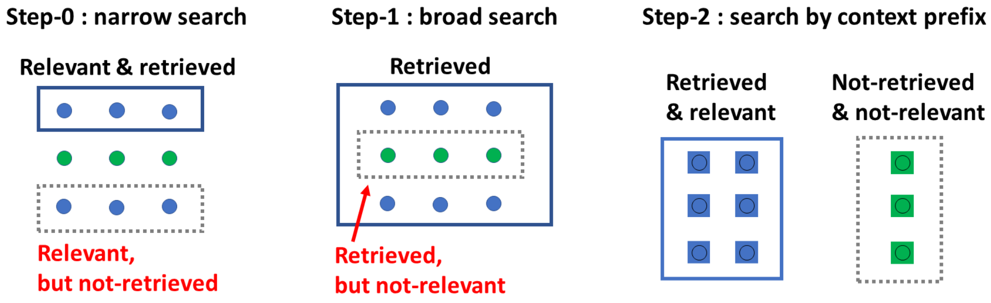


Figure 1: Context-sensitive keyword classification

the interests and curiosities of the researchers. For example, one keyword may appear in various research fields and be used frequently by many researchers, while another keyword may be used by only one researcher. In addition, the way in which trending keywords are used changes over time. In the past, the keyword “deep learning” was used to describe research on information technology. In recent years, however, “deep learning” has been used in all areas of research. Consequently, it is no longer possible to identify what the research is about with the keyword “deep learning” alone. It is necessary to confirm the meaning of the words by using the word context in the reports where the keywords are listed. Based on the above research background, our study in this report proposes a method of semantic disambiguation by extending the keywords in grant reports by the categories of the research fields in the KANENHI [1] review section table [2] (specifically, alphabet letters A-K).

There are previous studies ([3], [4], [5], [6]) that have conducted text analysis in the field of institutional research (IR). Our study makes a research contribution that differs from previous studies in that it specifically focuses on the polysemy of keywords in research grant reports and evaluates the proposed method for textual analysis by using actual data.

2 Method

Figure 1 illustrates the idea underlying our proposed method. The details of the method are described below.

- When the scope of keyword search is narrowed down from a large set of keywords to obtain keywords of interest, i.e., highly relevant keywords, in IR tasks, a small number of relevant keywords are included in the search results. Other relevant keywords are excluded from the search results. (Figure 1, step-0)

- When the range of search is widened, many relevant keywords are included in the search results, thus alleviating the failure of the search. However, due to ambiguity in the meaning of keywords, not-relevant keywords are also included in the search results. (Figure 1, step-1)

- To deal with word polysemy, a prefix is added to each word to distinguish the context in which the word appears so that only relevant keywords are retrieved, and not-relevant keywords are excluded from the search results. (Figure 1, step-2)

Table 1: Text-preprocessing for scientific keywords

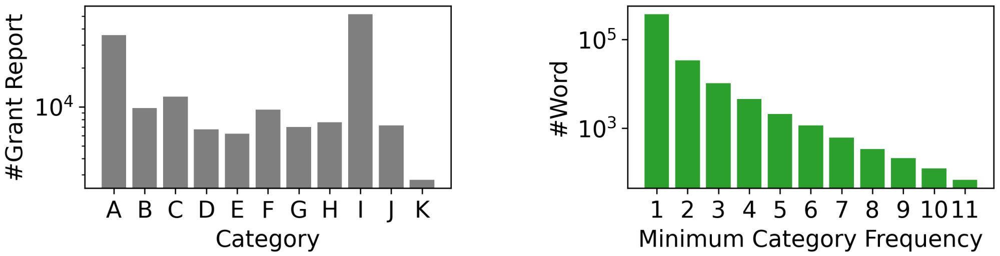
Grant ID	(1) Baseline	(2) Proposed	(3) Segmentation	(4) Hashtag
KKH-A-01	xxx yyy zzz	Axxx Ayyy Azzz	xx x yy y zz z	♯A xxx yyy zzz
KKH-A-02	xxx yyy zzz	Axxx Ayyy Azzz	xx x yy y zz z	♯A xxx yyy zzz
KKH-B-01	xxx yyy zzz	Bxxx Byyy Bzzz	xx x yy y zz z	♯B xxx yyy zzz
KKH-B-02	xxx yyy zzz	Bxxx Byyy Bzzz	xx x yy y zz z	♯B xxx yyy zzz
KKH-C-01	xxx yyy zzz	Cxxx Cyyy Czzz	xx x yy y zz z	♯C xxx yyy zzz
KKH-C-02	xxx yyy zzz	Cxxx Cyyy Czzz	xx x yy y zz z	♯C xxx yyy zzz

The prefixes attached to keywords, which play an important role in our study, are explained next. Table 1 presents a simple and concrete example of the prefixes for qualifying keywords. In this example, there are two adopted projects in each of the three categories A, B, and C, each representing a research field (e.g., A for Philosophy, B for Algebra, and C for Mechanics), with three keywords xxx, yyy, zzz representing research concepts (e.g., *sinsōgakushū* meaning Deep Learning, *sūrimoderu* meaning Mathematical Modeling, and *konchū* meaning Insects) mapped to them. As the three keywords appear in all three categories, the three categories cannot be distinguished by the keywords. (Table 1, (1) Baseline)

Therefore, in order to differentiate keywords in the categories in which they appear in the proposed method, a prefix indicating the category (e.g., A, B, or C) is added to the beginning of each word. The prefix for differentiating the context in which the word appears can be any character not included in the original word to prevent conflicts between words extended with the prefix and words not extended with the prefix. In our example, the prefix is safely added using the uppercase letters A, B, and C. Note that these letters do not overlap with x, y, or z for representing the keywords. (e.g., A*sinsōgakushū*, B*sinsōgakushū*, and C*sinsōgakushū*) This keyword extension gives us a total of 9 tokens for keywords, allowing us to differentiate keywords extended with prefixes as keywords with different notations across categories. (Table 1, (2) Proposed)

In text processing, when the number of characters in a string is large, the string is generally too specific to match each other. Therefore, the string is segmented to reduce specificity and make it easier for words to match each other. (Table 1, (3) Segmentation) In this study, the ambiguity of keywords matching each other too much is a problem to be solved. Hence, the fundamental idea in our study is not how to reduce the number of characters, but how to increase the number of characters by adding a prefix. Note that the process used in this study dit not involve the addition of information to the group of letters xxx, yyy, zzz. (Table 1, (4) Hashtag) It should be emphasized that our method is the "prefix addition" process, which adds a category modifier to make each token more identifiable. Hence, our method enables a search that asks what Grant ID contains the same keyword in the grant report in the same research field. For example, as both KKH-A-01 and KKH-A-02 contains Axxx, Ayyy, and Azzz, they are directly associated in keyword searches. In the actual data, Axxx, Ayyy, and Azzz may be A*sinsōgakushū*, A*sūrimoderu*, and A*konchū*, respectively. [1]

[1] Note that the proposed method is neither AND search in databases nor regular expression in text search. While Axxx specifies that the letter A precedes the letter sequence xxx and it can be used for retrieving only "Axxx," "A & xxx" may retrieve both "A xxx B yyy" and "B xxx A yyy" as the order of the letter sequences is not specified by the AND operator (&). In text search using regular expressions, a search pattern "A*xxx" may



(a) Number of grant reports by category

(b) Number of keywords by frequency

Figure 2: Keyword occurrences in grant reports (FY 2018-2023)

Table 1 represents a simplified example to illustrate the theoretical advantages of our proposed method. In the actual noisy data, there are many variations and combinations of word notations. Furthermore, the frequency of word occurrences is a mixture of density and sparseness. It is necessary to confirm by data experiments whether the proposed method is effective in realistic application.

3 Data

This study used data from XML files downloaded from the KAKEN [1] database on March 28, 2024. The collected data contained 1,031,131 grant documents. Some of the older grant documents are several decades old and do not have keywords listed. The number of grant documents with keywords listed was 811,395. The categories representing the research areas are redefined every few years, and the latest categories in the review section table [2] are not assigned to older grant documents. To be more specific,160,773 documents from FY2018 to FY2023 were assigned the latest review section. The number of approved proposals was almost evenly distributed each year as shown in Table 2. On the other hand, there was a large gap among the documents in the research fields, as shown in Figure 2 (a), with the number of K proposals being the lowest than the others, at 2,729. Figure 2 (b) shows the number of different words corresponding to the lowest frequency of occurrence in the categories. The number of words with a minimum frequency of 1 was 371,637. These words occur at least once in any category. The number of words with a minimum frequency of 2 or more was 34,345. These words appeared in more than one category and were considered to be ambiguous. The number of the most ambiguous words that appeared in all 11 categories was 69. These words included “deep learning,” “mathematical models,” “microorganisms,” “insects,” “cancer,” “led,” “gel,” “stability,” “gene expression,” and “enzymes.”[2]

4 Experiment

We conducted experiments to confirm the effectiveness of the proposed method. The effectiveness was measured by the accuracy in automatic document classification. Automatic

retrieve both “A xxx B yyy” and “A yyy B xxx” as the proximity of letters are not specified by the pattern.

[2] In the actual data, the corresponding Japanese words for these translated English words were as follows: *sinsōgakushū*, *sūrimoderu*, *biseibutsu*, *konchū*, *gan*, *LED*, *geru*, *anteisei*, *idenshihatsugen*, and *kōso*.

Table 2: Number of documents by year

	FY2018	FY2019	FY2020	FY2021	FY2022	FY2023	Total
#Document	23,795	28,964	28,598	25,882	28,246	25,288	160,773

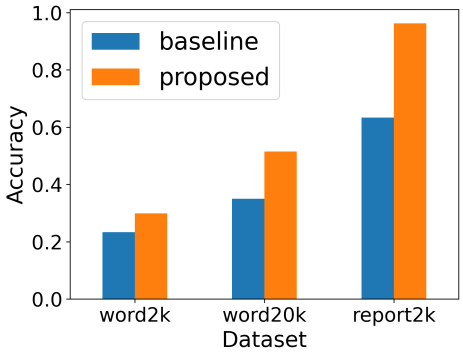


Figure 3: Comparison of accuracies

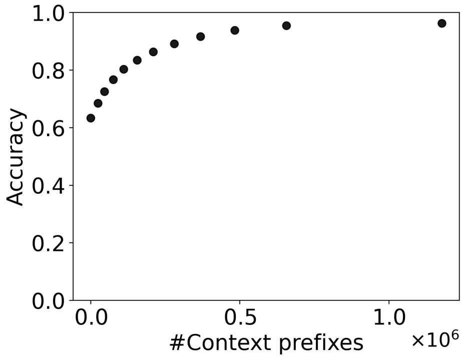


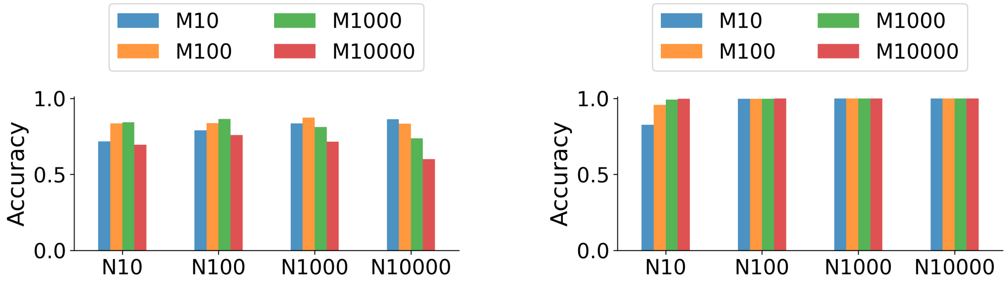
Figure 4: Accuracy by number of prefixes

document classification is one of the most common tasks in text analysis. Since our target data in this study was already assigned category labels for applying automatic document classification, it is desirable that costly text annotation by human experts was not required. In this study, the algorithm implementation for document classification was an SVM classifier in sklearn [7].

First, we conducted an experiment to compare the accuracies of the baseline and proposed methods using the data with the latest review section categories from FY2018 to FY2023. Specifically, we prepared three different datasets for comparison. They are (1) a dataset with 2,000 single words for each category (referred to as word2k), (2) 20,000 single words (referred to as word20k), and (3) 2,000 groups of words per grant report (referred to as report2k). Text preprocessing was performed on each of the datasets using the baseline and proposed methods.

The obtained results are shown in Figure 3. As can be seen in the figure, the accuracy of the proposed method shown in orange exceeds the accuracy of the baseline method shown in blue. In addition, word2k, which is decomposed into individual words, loses the contextual relations between words in the keyword lists, and the accuracy of the baseline method was remarkably low. Applying the proposed method to this dataset did not significantly improve the accuracy. On the other hand, word20k, with a larger amount of data than word2k and a richer amount of information about the frequency of words, yielded higher accuracies than word2k for both the baseline and proposed methods. However, this dataset also had limited accuracy improvement due to the loss of relationships between words in the keyword lists. For groups of words per research report that contained word context information, the baseline method achieved an accuracy of 0.633, while the proposed method improved the accuracy to 0.962 by adding contextual prefixes.

To investigate the difference in accuracy between the baseline and proposed methods in more detail, we conducted an experiment in which the number of words to be prefixed



(a) Baseline

(b) Proposed

Figure 5: Accuracies dependent on M and N values

was varied according to the category frequency of the words in Figure 2 (b). The results are shown in Figure 4. In this scatter plot, the point on the leftmost X-axis corresponding to a value of 0 has no context-aware prefixes, which is equivalent to the baseline method. The point on the rightmost X-axis corresponding to the highest value, 1,177,824, is the case where all words were prefixed and the proposed method was applied to words with a category frequency of 1 or higher. The points in the middle correspond to the minimum category frequency of 2 to 11. It was confirmed that as the information representing the contextual relationship between words increases, the accuracy increases gradually, and once all words are prefixed the accuracy reached the maximum value.

Next, we conducted an experiment to compare the accuracy of automatic document classification between the baseline and proposed methods by creating a dataset whose accuracy was controlled by two search parameters, N and M proposed in a previous study [3]. For the experiments, 16 experimental datasets were created by varying the values of N and M to 10, 100, 1,000, and 10,000, respectively, and experiments on automatic document classification were conducted. The results are shown in Figure 5.

As can be seen from the figure, the proposed method outperformed the baseline method in all datasets. It is noteworthy when the words N and M are large, the baseline method performed poorly due to the ambiguity of the word meanings. Specifically, the larger the data size, the worse the accuracy becomes with the baseline method. In contrast, the proposed method marked accuracies of approximately 1.000, which confirmed that the idea of our proposed method was effective.

5 Discussion

Our proposed method is an outcome of trial and error using various techniques for text analysis. In addition to the experiments described above, we also conducted experiments on the word segmentation and the hashtag that were described in Section 2. As a result, these techniques were not effective; rather, a slight degradation in performance was observed. Research management tasks in IR require diverse expertise from data analysis to university management[8]. Our study in this report proposes a novel method from the viewpoint of word sense disambiguation ([9], [10]). As our method has been validated through experiments with open data that can be used for research management tasks, our study is expected to provide new insights to the IR community.

Keywords written in grant reports may be fully understood by each of the researchers who wrote the report. However, such keywords are often recognized differently by researchers in different research fields. When a document for submitting to the government, writing a document objectively without ambiguity is necessary. Thus, there is a special need among IR professionals to effectively utilize advanced expertise accumulated in large text databases. Our method resolves semantic ambiguity when searching and classifying keywords representing research topics; thus, fulfilling such requirements in text analysis for IR tasks.

To verify the feasibility of our method, we confirmed the processing times for both the baseline and proposed methods. The processing times for the baseline is the sum of the following three operations:

- Downloading all XML data, parsing, and storage the data into databases ($T1$)
- Concatenating stored keyword strings in the database ($T2$)
- Training and testing models using machine learning ($T3$)

On a desktop PC (Ubuntu 20.04, Intel Core i9-9820X CPU @ 3.30GHz, 256 GB RAM), $T1$, $T2$, and $T3$ took approximately 1 day, 1 second, and 1 minute, respectively. The baseline method involves string aggregation, whereas the proposed method involves string concatenation and aggregation.[3] Hence, the time taken for the proposed method ($T2_p$) is longer than that for the baseline ($T2_b$). Specifically, $T2_b$ was 1.0067 second, and $T2_p$ was 1.0461 second. The additional processing time of 0.0394 second required by the proposed method is sufficiently short in comparison with the total processing time of $T1+T2+T3$. It demonstrates that the proposed method is feasible in realistic IR tasks.

6 Conclusion

Our study proposes a method for word sense disambiguation in the reports of accepted research grants. The proposed method adds a prefix indicating the research field to each word. The added prefix can identify the meaning of a keyword considering the context in which the keyword appears.

The experiments on real data confirmed that the proposed method outperformed the baseline method on small datasets from FY2018 to FY2023 with the latest review section categories. We also confirmed that the proposed method outperformed the baseline method on larger datasets constructed by automatically assigning review section categories. When the proposed method narrows down the keyword search and the scope of the search is too small, the coverage of word variations is low, limiting the improvement in accuracies of both the baseline and proposed methods. When the scope of the search is broad, and the data size is sufficiently large, the model trained by the machine learning algorithm is successful in covering a wide variety of keywords, and the accuracy obtained with our proposed method is nearly 1.000, confirming that a notably high level of effectiveness can be achieved.

Our findings in this study are expected to be useful and applicable in research management tasks in IR. We will consider applying the proposed method to other text data with word sense disambiguation in our future study.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number JP23K11764. We would like to thank the reviewers who carefully read our manuscript and provided useful comments.

[3] The details of these data operation is explained in the documentation of PostgreSQL[11].

References

[1] The National Institute of Informatics, “KAKEN: Grants-in-Aid for Scientific Research Database,” https://kaken.nii.ac.jp/ .

[2] Japan Society for the Promotion of Science, “The Review Section Table for Basic Sections (lists),” https://www-kaken.jsps.go.jp/kaken1/shoukubunListEn.do .

[3] M. Yasukawa and K. Yamazaki, “Entity Linking among Categorized Knowledge Resources for Computer Science Curricula,” *IIAI Letters on Institutional Research*, vol. 3, no. LIR152, pp. 1–14, 2023.

[4] T. Tsumagari, N. Nakazato, and T. Tsumagari, “Student’ Interests and Career Understanding: A Topic Analysis of First-year Career Courses,” *IIAI Letters on Institutional Research*, vol. 1, no. LIR013, pp. 1–8, 2022.

[5] A. Itoh, H. Ito, S. Matsumoto, I. Noda, K. Bannaka, K. Nishiyama, T. Kirimura, T. Kunisaki, K. Mitsunari, K. Murakami, R. Kozaki, A. Kishida, M. Kondo, S. Imai, M. Mori, Y. Nakata, M. Omori, and K. Takamatsu, “A Two-Step Approach for Syllabus Development and Evaluation using Machine Learning such as Doc2Vec based on Eduinformatics,” *IIAI Letters on Institutional Research*, vol. 3, no. LIR128, pp. 1–11, 2023.

[6] M. Yasukawa and K. Yamazaki, “Detecting Transition of Research Themes using Time-oriented Attributes in Governmental Funding,” *International Journal of Institutional Research and Management*, vol. 7, no. 1, pp. 1–17, 2023.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] K. Robershaw and B. Wolf, “Research analytics: A systematic literature review,” *Social Science Research Network (SSRN)*, 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4363262

[9] M. Abeysiriwardana and D. Sumanathilaka, “A survey on lexical ambiguity detection and word sense disambiguation,” *IEEE International Colloquium on Signal Processing and its Applications (CSPA 2024)*, 2024. [Online]. Available: https://arxiv.org/abs/2403.16129

[10] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, “Recent trends in word sense disambiguation: A survey,” in *International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021, pp. 4330–4338. [Online]. Available: https://www.ijcai.org/proceedings/2021/0593.pdf

[11] The PostgreSQL Global Development Group, “Chapter 9. Functions and Operators,” https://www.postgresql.org/docs/current/functions.html .