Towards Extracting Formulaic Expressions from Japanese Scholarly Papers

Kenichi Iwatsuki[1][0000-0002-3428-2953]

[1] The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, Japan
iwatsuki@iktta.org

Abstract. While scholarly papers in many disciplines are written in English, non-English papers have been published. Formulaic expressions used in research articles have been studied, but past work mainly focused on English formulaic expressions. In this study, we applied an existing formulaic expression extraction method that was originally proposed for English papers to introduction sections of Japanese papers on natural language processing. The results show that the extraction is to some extent successful. However, the paucity of dataset of scholarly papers hinders the construction of a comprehensive list of formulaic expressions and comparison among multiple disciplines.

Keywords: Formulaic expressions, Scholarly papers, Japanese language.

1 Introduction

English is regarded as the lingua franca of scholarly papers in many disciplines. While most papers have been published in English recently, non-English articles are still published. In Japan, the Japanese language is mainly used in some disciplines including jurisprudence. Even in computer sciences, many domestic conferences and journals accept papers written in Japanese. Some research topics are specific to situations and industry in Japan or the Japanese language and literature.

The intranational articles provide useful information for specific communities; thus, it is important to develop technologies for processing domestic scholarly papers as well as English papers. Scientific document processing is a field related to natural language processing, information retrieval, and digital libraries, and its importance has been increased because the number of scholarly articles has been growing. Recent development of the technologies is remarkable. SciBERT, a BERT-based language model pre-trained on scholarly articles of biomedicine and artificial intelligence was released in 2019 [1]. A large dataset for citation intention classification was published in 2019 [2]. SciERC [3] is a new dataset for multitask learning for the extraction of scientific entities and their relations. The progress came from large datasets; thus, non-English literature is not to benefit from the state-of-the-art technologies.

Academic writing is another perspective of scholarly papers. The genre of scientific papers has its own style of writing. Thus, how a paper is written should be learnt even by native speakers. Using formulaic expressions is helpful in writing scientific papers.

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 74–82.

Formulaic expressions assure that the wordings are grammatical and conform to conventions of a specific community. To make the most of them, a list of formulaic expression should be created. Although there are many studies on formulaic expressions used in English papers, less attention has been paid to Japanese academic formulaic expressions.

A problem lying in formulaic expression research is how to extract formulaic expressions from a corpus. Many past studies extracted word *n*-grams called *lexical bundles*, but it was pointed out that the *n*-grams are not always helpful for pedagogical purposes [4].

Although methods for extracting formulaic expressions from scientific papers written in English were proposed, it is still unclear whether the methods can be applied to those written in another language. Thus, in this study, we applied an existing formulaic expression extraction method [5] to the introduction sections of Japanese research papers.

The results show that the method extracted formulaic expressions successfully, but most of them contained the end of sentences because the method utilised a sentential root that comes near the end of a sentence in Japanese, a SOV language.

We argue that the scarcity of machine-readable dataset of Japanese scholarly papers hinders comparison among disciplines or sections. Constructing the dataset is an urgent task for advancement of the research on Japanese academic formulaic expressions.

2 Related Work

2.1 Formulaic Expressions in Academic Prose

Formulaic expressions are often classified based on communicative functions. Each component of scholarly papers has its communicative function such as *showing the aim of the paper* and *describing the results*. The combinations of formulaic expressions and communicative functions [7] are useful for academic writing assistance [8]. Automated construction of a list of formulaic expressions with communicative function labels was proposed [9].

Formulaic expressions have been reported to be specific to disciplines. For example, '*was stirred at room temperature for*' appears in chemistry papers while '*study was carried out in accordance with the*' occurs in psychology papers [9]. Grammatical difference has been also observed; active voice is preferred such as '*we present a new*' in computational linguistics papers, but passive voice is selected such as '*was reported in*' in chemistry papers [9].

2.2 Scholarly Paper Processing

Natural language processing in scientific papers attracts more and more attentions. However, recent progress on this topic have been made only on limited disciplines. SciBERT [1] looks as if it had addressed many disciplines in science, but actually it was trained only on biomedical and artificial intelligence papers. ScispaCy [10] is a

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 74–82.

model for spaCy aiming at processing biomedical text, not comprehensive scientific text. Therefore, even in English language, only a small part of science has been addressed recently.

2.3 State of the Japanese Language Processing

The number of native speakers of Japanese is more than 100 million, almost all of which live in Japan. Thus, most work on natural language processing for the language was conducted by Japanese research institutions.

Kyoto University Text Corpus [12] and the balanced corpus of contemporary written Japanese (BCCWJ) are famous corpora of Japanese [13]. Morphological analysis is an important part of processing Japanese because words are not separated. Several analysers have been presented. JUMAN [14], ChaSen (Markov model) [15], MeCab (conditional random fields) [16], JUMAN++ (recurrent neural network) [17], and Sudachi [18] are free morphological analysers.

Pre-trained models for BERT have been provided by multiple laboratories. Most of the models were trained on Japanese Wikipedia [19–21], while models trained on a news corpus [22], web pages written in Japanese [23] are available.

3 Methods

3.1 Corpus

Unlike English scientific papers, for which datasets comprising computer-readable text are available, it is difficult to obtain natural language text extracted from Japanese scientific papers. Although a repository called J-STAGE run by Japan Science and Technology Agency collected millions of scholarly papers, most of them are not open-access and even open-access articles are formatted in the portable document format, which demands laborious processing[1].

Avoiding the complexity, we used a corpus that consisted of LaTeX sources of journal articles on natural language processing (Journal of Natural Language Processing). The corpus was provided by the Association for Natural Language Processing and was available on the web[2].

The journal accepts manuscripts written in both Japanese and English, so that we picked only Japanese articles out of them. The languages were detected by the document class identifier of the sources.

[1] Several tools exist for extracting text from PDF files. The pdftotext and poppler extract text from PDFs. Figures, tables, page numbers, and multicolumn layout should be addressed properly.

[2] https://www.anlp.jp/resource/journal_latex/index.html

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 74–82.

3.2 Pre-Processing

The pre-processing was three-fold: extraction of narrative text, identification of sections, and sentence splitting. In the final dataset, Japanese sentences without any LaTeX commands were assigned section labels.

In the first step, all LaTeX commands were removed. Additionally, tables, figures, lists, and equations were removed. In-line mathematical expressions such as x were replaced with a special token *MATH*.

The section identification was conducted using section headings. The section headings varied so significantly that the extraction was performed only in the introduction sections of the corpus. The headings 'はじめに' and '序論' were integrated into the introduction.

The sentence splitting and tokenisation[3] were conducted with spaCy[4] and GiNZA[5] [11], a library for processing the Japanese language on spaCy.

The number of words and sentences in the final dataset is listed in Table 1.

Table 1. Numbers of words and sentences in the final dataset.

Sentences	Words
6,454	326,824

3.3 Extracting Formulaic Expressions

There are two approaches for the methodology for extracting formulaic expressions: the corpus-level and sentence-level approaches [5]. With the corpus-level approach, word *n*-grams are first extracted, and then the *n*-grams are filtered based on frequency, mutual information, or other metrics. With the sentence-level approach, one formulaic expression is extracted from one sentence; thus, this approach is regarded as a sequence labelling problem.

The corpus-level approach requires determining lengths of targeted *n*-grams in advance. Moreover, it causes a *span problem* [6], difficulty in determining which extracted *n*-grams should remain or be discarded, e.g. '*in this paper we propose*' and '*paper we propose a new*', both of which are 5-grams and the latter is not a good formulaic expression.

The sentence-level approach avoids these problems because only one formulaic expression is extracted from one sentence. Thus, we adopted this approach in this study. The limitation of this approach is that it is impossible to extract more than one formulaic expression from a sentence.

---

[3] Different from English, words are not separated by white spaces in Japanese; thus, the tokenisation includes not only identification of words' part-of-speech but also word segmentation.

[4] https://spacy.io

[5] https://megagonlabs.github.io/ginza/

We adopted the sentence-level formulaic expression extraction method proposed in the previous work [5]. The extraction method consists of two steps: named entity removal and *n*-gram extraction.

In the original settings, scientific entities including names of bacteria and datasets and named entities such as '*Helsinki*' were removed because the entities were specific to details of a research article and did not compose formulaic expressions. Although scientific entity recognition is one of the popular tasks in processing scientific documents written in English, datasets or methods have not been developed for Japanese scholarly papers. Thus, we utilised only the named entity recognition provided by the GiNZA and removed them from the sentences.

In the second step, two kinds of *n*-grams are extracted.

1. longest *n*-grams satisfying a frequency threshold and
2. longest *n*-grams satisfying the frequency threshold and containing a sentential root.

Afterwards, the two *n*-grams are combined to be a resulting formulaic expression. If the two *n*-grams are not adjacent, a slot '*' is inserted in between.

The sentential root was identified using spaCy and GiNZA. Although in the original setting, the frequency threshold was set to three times in a corpus, we mitigated the threshold to twice in a corpus because the size of the corpus was much smaller than that used in the previous work.

For the sake of comparison, we also extracted word *n*-grams ($n \geq 3$), which is regarded as a corpus-level approach.

4 Results

Table 2 lists the top-10 frequent formulaic expressions extracted with the sentence-level method and word *n*-grams extracted based on frequency. All the top-10 formulaic expressions contained the punctuation mark that denotes the end of a sentence. The frequent word *n*-grams are shorter than the formulaic expressions because the shorter the *n*-grams are, the more frequently they occur.

Of course the formulaic expressions were also extracted by the corpus-level method, but it is difficult to select them from all the *n*-grams. While the only formulaic expression occurred 12 times in the corpus was 'が提案されている．' ('*has been proposed.*'), 401 *n*-grams that occurred 12 times in the corpus were extracted and many of them were non-sense. The number of the extracted formulaic expressions is 457 while the number of word *n*-grams occurring at least twice in the corpus is 118,217.

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 74–82.

Table 2. Extracted formulaic expressions and number of occurrences (Freq.) in the corpus.

Sentence-level method		Corpus-level method	
Formulaic expressions	Freq.	Word *n*-grams	Freq.
が提案されている.	12	として	1,360
が必要である.	11	ている.	1,346
という問題がある.	8	では,	1,247
となっている.	8	である.	1,074
について説明する.	7	れている	880
である.	7	されて	865
を行っている.	7	について	861
と呼ぶことにする.	6	において	702
手法が提案されている.	6	している	641
とされている.	6	されている	578
する必要がある.	6		
がなされている.	6		

5 Discussion and Conclusion

The top-10 frequent formulaic expressions contained the punctuation mark denoting the end of a sentence. Thus, the extraction method is likely to extract formulaic expressions appearing near the end. This phenomenon happened because the sentential root was used to identify formulaic expressions. The difference of word orders between English and Japanese caused this.

However, the meanings and functions of the extracted formulaic expressions are not much different from English ones. Table 3 lists the extracted formulaic expressions and functions. The functions, referring to past work, problem statement, definition of terms, providing background, and outline of the paper are used in English scientific papers.

Formulaic expressions are specific to disciplines. Some formulaic expressions occur many times in computer science while others are used mainly in chemistry but not in computer science. Many studies have been conducted to analyse discipline- and section-specific formulaic expressions. To construct a multi-disciplinary list of academic formulaic expressions, corpora of scholarly articles must be prepared. We argue that the keys to the construction are formats of articles and open access to full text. Formats such as HTML, XML, and plain text are easy to process. For instance, PubMed Central provides XML data of biomedical papers. ArXiv provides LaTeX sources of its collection.

Table 3. Functions of extracted Japanese formulaic expressions.

Functions	Formulaic Expressions (English translations)	Freq.

Referring to past work	が提案されている．(has been proposed)	12
	手法が提案されている．(method is proposed)	6
	を提案している．(has proposed)	5
	する手法を提案している．(has proposed a method to)	5
	手法が数多く提案されている．(many methods have been proposed)	5
Problem statement	が必要である．(is necessary)	11
	という問題がある．(there is a problem that)	8
	する必要がある．(it is necessary to)	6
Definition of terms	と呼ぶことにする．(decided to refer to … as)	6
Providing background	が注目されている．(is paid attention to)	5
	の研究が盛んである．(studies on … are popular)	4
	が盛んに研究されている．(has been studied actively)	4
Outline of the paper	本論文の構成は以下の通りである．(the structure of this paper is as follows.)	5

References

1. Beltazy, I., Lo, K., Cohan, A.: SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620. Association for Computational Linguistics, Hong Kong (2019).
2. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural Scaffolds for Citation Intent Classification in Scientific Publications. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3586–3596. Association for Computational Linguistics, Minneapolis (2019).
3. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3219–3232. Association for Computational Linguistics, Brussels (2018).
4. Swales, J. M.: The futures of EAP genre studies: A personal viewpoint. Journal of English for Specific Purposes 38, 75–82 (2019).
5. Iwatsuki, K., Aizawa, A.: Extraction of Formulaic Expressions from Scientific Papers. In: The AAAI-21 Workshop on Scientific Document Understanding. Online (2021).
6. Iwatsuki, K., Aizawa, A.: Using Formulaic Expressions in Writing Assistance Systems. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2678–2689. International Committee on Computational Linguistics, Santa Fe (2018).
7. Cortes, V.: The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for Academic Purposes 12, 33–43 (2013).

8. Mizumoto, A., Hamatani, S., Imao, Y.: Applying the Bundle–Move Connection Approach to the Development of an Online Writing Support Tool for Research Articles. Language Learning 67(4), 885–921 (2017).
9. Iwatsuki, K., Aizawa, A.: Communicative-Function-Based Sentence Classification for Construction of an Academic Formulaic Expression Database. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. to appear (2021).
10. Neumann, M., King, D., Beltazy, I., Ammar, W.: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. Association for Computational Linguistics, Florence (2019).
11. 松田寛, 大村舞, 浅原正幸: 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. In: 言語処理学会第25回年次大会発表論文集, pp. 201–204. Association for Natural Language Processing, Nagoya (2019).
12. Kurohashi, S., Nagao, M.: Building a Japanese Parsed Corpus while Improving the Parsing System. In: Proceedings of the 1st International Conference on Language Resources & Evaluation, pp. 719–724. European Language Resources Association, Granada (1998).
13. Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., Den, Y.: Balanced corpus of contemporary written Japanese. Language Resources & Evaluation 48, 345–371 (2014).
14. Myoki, Y., Matsumoto, Y., Nagao, M.: User Customizable Japanese Dictionary System and Morphological Analyzer. In: 全国大会講演論文集, pp. 17–18. Information Processing Society of Japan, Hachioji (1991).
15. Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y.: Japanese Morphological Analysis System ChaSen version 2.0 Manual. Information Science Technical Report TR99009 (1999).
16. Kudo, T., Yamamoto, K., matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230–237. Association for Computational Linguistics, Barcelona (2004).
17. Morita, H., Kawahara, D., Kurohashi, S.: Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2292–2297. Association for Computational Linguistics, Lisbon (2015).
18. Takaoka, K., Hisamoto, S., Kawahara, N., Sakamoto, M., Uchida, Y., Matsumoto, Y.: Sudachi: a Japanese Tokenizer for Business. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 2246–2249. European Language Resources Association, Miyazaki (2018).
19. 柴田知秀, 河原大輔, 黒橋禎夫: BERTによる日本語構文解析の精度向上. In: 言語処理学会第25回年次大会発表論文集, pp. 205 -- 208. Association for Natural Language Processing, Nagoya (2019).
20. Kikuta, Y.: BERT Pretrained model Trained On Japanese Wikipedia Articles. https://github.com/yoheikikuta/bert-japanese (2019).
21. National Institute of Information and Communications Technology: NICT BERT 日本語 Pre-trained モデル. https://alaginrc.nict.go.jp/nict-bert/index.html (2020).
22. Stockmark Inc.: 大規模日本語ビジネスニュースコーパスを学習したBERT事前学習済（MeCab利用）モデルの紹介. https://qiita.com/mkt3/items/3c1278339ff1bcc0187f (2019).

23. Zhao, X., Hamamoto, M., Fujihara, H.: Laboro BERT Japanese: Japanese BERT Pre-Trained with Web-Corpus. https://github.com/laboroai/Laboro-BERT-Japanese (2020).