Extraction of Formulaic Expressions from Scientific Papers

Kenichi Iwatsuki,[1] **Akiko Aizawa**[2,1]

[1] The University of Tokyo, Japan
[2] National Institute of Informatics, Japan
iwatsuki@nii.ac.jp, aizawa@nii.ac.jp

Abstract

Phrasal patterns, such as ‘in this paper we propose’, are often used in scientific papers. These are called *formulaic expressions* (FEs) and constitute sentential communicative functions (CFs) that convey how a sentence should be read by the readers. FEs are useful for scientific paper analyses and academic writing assistance, but FE extraction methods have thus far not been investigated in detail. In this paper, we propose a sentence-level FE extraction method in which the CFs are taken into account. The proposed method is compared to existing methods to demonstrate that it is better at CF-oriented FEs.

1 Introduction

In scientific papers, the authors often use several fixed phrasal patterns that are specific to the genre, such as ‘in this paper, we propose’. These patterns are called *formulaic expressions* (FEs) or *formulaic sequences*. FEs convey the intentions of the authors to the readers, i.e., the manner in which a sentence should be understood. This characteristic of the FE is called *communicative function* (CF). For example, the phrase ‘in this paper, we propose’ conveys the CF of the sentence meaning ‘showing the aim of the paper’. FEs are useful for understanding the composition of a scientific paper and are helpful in writing the paper.

A few studies have been reported on addressing the extraction of FEs and subsequent assignment of CF labels to them (Cortes 2013; Mizumoto, Hamatani, and Imao 2017). However, these works have not rigorously investigated whether the extracted FEs convey the CFs of a sentence. Extracting word n-grams with frequency thresholds has been reported in several studies, although frequent FEs do not always convey the sentential CFs. Machine-learning approaches have hitherto been scarcely adopted because of the dearth of sufficient FE-annotated resources.

In this paper, we propose a new sentence-level FE extraction method and compare it to several existing methods. We assume that a single FE is extracted from each sentence because it conveys the entirety of the CF of that sentence. The proposed method consists of two steps. First, the named and scientific entities are removed from the sentence. Second, two types of n-grams are extracted from the sentence.

Then, the extracted FEs are evaluated based on whether they convey the sentential CFs. The results of manual evaluations show that the proposed method can extract more FEs representing the CFs of sentences than existing methods.

Considering the compilation of a list of FEs, which will be a possible application of the FE extraction, removing noisy FEs and enhancing precision is important. Thus, we test how effective filtering FEs based on the number of occurrence of an FE is, and show that it improves precision much.

2 Datasets

We used a CF-labelled sentence datasets that were made from scientific papers of four disciplines: computational linguistics (CL), chemistry (chem), oncology (onc), and psychology (psy). Each discipline consists of four sections; introduction, methods, results, and discussion; thus, 16 datasets were used (combination of four disciplines and four sections). The numbers of sentences and words in these datasets are listed in Table 1. Compared with some of the existing studies, in which the sizes of the corpora were around 2 million (Simpson-Vlach and Ellis 2010) or 8 million (Mizumoto, Hamatani, and Imao 2017) words, we determined that the datasets are sufficient.

3 Methods

3.1 Two Approaches in FE Extraction

Two main approaches were considered here for extracting the FEs: corpus- and sentence-level approaches. In the corpus-level approach, the FEs are extracted from the entire corpus, whereas in the sentence-level approach, a single FE is extracted from each sentence (Figure 1). The corpus-level approach may cause problems with deciding the FE size and overlap between FEs (Iwatsuki and Aizawa 2018). For example, when 4-grams are extracted in the experiments, the phrases ‘paper we propose a’ and ‘we propose a method’ were both extracted, but it is difficult to determine which of these is a better FE. In contrast, the sentence-level approach is free of this problem because it does not have a fixed length for the n-gram. Since a single FE is extracted from each sentence, only ‘in this paper we propose a method’ is extracted. Therefore, we adopt the sentence-level approach in

Discipline	Section	Sentences	Words
CL	introduction	266,904	5,934,772
	methods	362,477	7,469,502
	results	507,592	10,176,904
	discussion	111,052	2,481,983
Chem	introduction	285,810	7,526,537
	methods	376,583	8,655,414
	results	721,960	18,308,473
	discussion	175,266	4,443,967
Onc	introduction	441,141	11,051,210
	methods	976,205	20,615,171
	results	1,069,044	27,059,136
	discussion	834,641	20,897,907
Psy	introduction	484,615	13,944,874
	methods	429,155	9,898,714
	results	288,754	7,756,912
	discussion	453,118	12,641,250

Table 1: Numbers of sentences and words in each discipline and section in the prepared CF-labelled sentence datasets.

the remaining experiments. We compared two corpus-level and two sentence-level methods with the proposed method.

3.2 Corpus-Level Extraction

Frequent N-grams Word n-grams were extracted from the dataset, and depending on the frequency-based threshold, the infrequent FEs were removed. Although various studies have used different lengths and frequency thresholds for the n-grams, we extracted FEs whose lengths were three words or greater, and followed the method in Cortes (2013) for the frequency thresholds: 20 per million words (pmw) for four-word or shorter n-grams, 10 pmw for five-word phrases, 8 for six- and seven-word phrases, and 6 pmw for phrases longer than seven words.

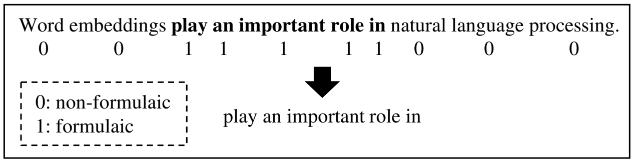
Word embeddings **play an important role in** natural language processing.
0 0 1 1 1 1 1 0 0 0
0: non-formulaic
1: formulaic
play an important role in

Figure 1: Sentence-level FE extraction.

Lattice FS This approach was originally proposed by Brooke, Šnajder, and Baldwin (2017), where n-grams are first extracted and later selected based on the concepts of *covering*, *clearing*, and *overlap*. Covering indicates that if the number of instances of 'we propose' is almost the same as those of 'we propose a new', the longer FE would explain the presence of the shorter FE. Clearing indicates the opposite idea to covering. Overlap indicates that the expressions 'in this paper we' and 'this paper we proposed' should not be accepted as FEs at the same time. These three concepts are expressed in mathematical form, and the FEs are optimised computationally. We used an implementation available[1].

3.3 Sentence-Level Extraction

Frequency-Based Filtering Based on the frequency, each word of a sentence is labelled as either formulaic or non-formulaic. Non-formulaic words are removed, and the remaining words are regarded as the FE. We used two frequency thresholds, namely 1/50,000 and 1/100,000 words.

LDA-Based Filtering Liu et al. (2016) proposed utilising latent Dirichlet allocations (LDA) because they assumed that topic-specific words do not comprise FEs. Thus, each word of a sentence was judged as either topic-specific or topic-independent based on the following criterion:

$$\mathrm{P}(w) = 1 - \frac{\max p_w(i)}{\sum p_w(i)},$$

where $p_w(i)$ is the probability of the word w in a topic i. If $P(w)$ is greater than the threshold, w is formulaic. We use $P(w) > 0.65$ and 10 topics, which was reported optimal.

Proposed Method The proposed method comprises two steps: (1) removing named and scientific entities and (2) extracting longest word n-grams (Figure 2). The first step was based on the idea that the named and scientific entities, including places, organisations, materials, and methods, such as 'Helsinki' and 'word embeddings', do not constitute FEs. In the second step, dependency parsing was applied to the sentences to determine their roots. After removing the named and scientific entities, two types of word n-grams were labelled as formulaic:

1. the longest word n-gram satisfying a frequency threshold;
2. the longest word n-gram that contains a root of the sentence and satisfies the frequency threshold.

If multiple FEs of the same lengths were found, the most frequent one was prioritised.

We focused on the longest word sequences because Cortes (2013) observed that lengthy FEs, such as 'the rest of the paper is organized as follows' existed. Additionally, we assumed that in several cases, sentential CFs were realised around the root of the sentence, so that two types of n-grams should be extracted. Specifically, n-grams whose lengths were less than three words were ignored because such FEs would be too short. The remaining words in the sentence after n-gram extraction were removed. The frequency threshold was thus set to 3 to collect the maximum number of FEs. The entity removal was conducted with ScispaCy (Neumann et al. 2019).

In the example in Figure 2, the root word is '*show*'. The longest n-gram satisfying the threshold and containing the root would thus be '*the results show that*', while '*is significantly better than*' would be another n-gram that does not contain the root. There could also be cases where these two types of FEs overlap or be the same.

[1] https://github.com/julianbrooke/LatticeFS

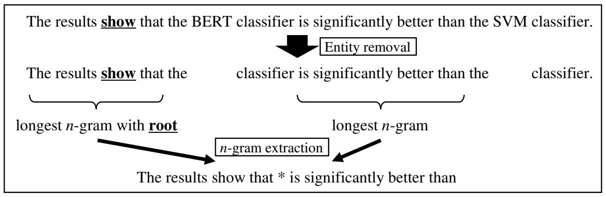
The results **show** that the BERT classifier is significantly better than the SVM classifier.
Entity removal
The results **show** that the classifier is significantly better than the classifier.
longest n-gram with **root**
longest n-gram
n-gram extraction
The results show that * is significantly better than

Figure 2: The proposed FE extraction method.

3.4 Filtering FEs

For compiling a list of FEs, which is one of the applications of the FE extraction, it is not always necessary to use all these FEs extracted from every sentence. It is more important to discard non-FEs. Because the word sequences that occur only once or twice are not formulaic, filtering FEs based on the number of the occurrence is effective. Therefore, we set thresholds of the number of FE occurrence in the dataset, and removed FEs not satisfying the thresholds.

4 Results

We randomly chose 100 sentences from the sentence dataset to evaluate the FE extraction. For the sentence-level methods, a single FE was extracted from each sentence. For the corpus-level methods, the FEs and sentences were not clearly connected. Thus, we randomly selected a single FE from the set of extracted FEs for each sentence.

The evaluations were then conducted manually. Three annotators were asked to check if the FEs extracted with each method had the same CFs as the sentences from which they were extracted and if these were reusable when writing scientific papers. The FEs were presented to the annotators simultaneously, and the method that was applied to the FE was not disclosed. A total of 100 combinations of sentences and FEs were randomly selected for evaluations.

The results of the evaluations are shown in Table 2, and the proposed method is observed to show clear advantage over other baselines in the FE extraction.

Method	$\geq$ 2/3	3/3	κ
Frequent n-grams	0.30	0.09	0.36
Lattice FS	0.07	0.03	0.30
Frequency-based (1/50,000)	0.04	0.02	-0.36
Frequency-based (1/100,000)	0.05	0.02	-0.39
LDA-based	0.08	0.03	-0.20
Proposed (Step 1)	0.13	0.05	-0.27
Proposed (Step 2)	0.54	0.28	0.23
Proposed (Step 1+2)	0.58	0.39	0.44

Table 2: Ratios of FEs that two or three out of the three ($\geq$2/3) and all three (3/3) annotators labelled as correct. Fleiss's kappa is also shown.

Table 3 shows the thresholds of the number of occurrence of FEs and scores. From the table, it can be seen that if FEs occurring less than three times in a corpus are ignored, the precision would change from 0.39 (39/100) to 0.49 (24/53). It should be noted that the recall cannot be calculated because there are no available FE-annotated resources.

Occurrence	≥ 1	≥ 3	≥ 5	≥ 7
Ratio of 3/3	0.28	0.45	0.55	0.53
#	39/100	24/53	21/47	21/46

Table 3: Ratios of FEs whose score was 3/3 and filtering thresholds of occurrence.

5 Discussion

5.1 Errors in Entity Recognition

We analysed the errors (FEs that 1/3 or less annotators judged as correct) in the proposed method. The errors in the entity recognition (step 1) accounts for approximately 60% of all the errors. They can be classified into two types: (1) entities are not removed and (2) formulaic words are removed as entities though they are not entities. Most of the errors were the type (2).

Table 4 lists the examples of this error. From this table, it can be seen that formulaic words such as '*table*' and '*investigated*', which are indispensable for representing the CFs, were removed. When formulaic words are removed at this stage, meaningful n-grams are not to be extracted in the step 2. This results infer that entity recognition is crucial to the proposed method and should be improved much.

5.2 Errors in N-grams

Another type of errors is the errors in the n-gram extraction (step 2). In the proposed method, we extracted two different n-grams: the longest n-gram containing the sentential root and the longest n-gram that does not necessarily contain the root, both of which satisfied the threshold of the number of occurrence in the corpora.

The majority of this error is that the extracted two n-grams are the same but do not contain CF-realising part. Table 5 lists the examples of this error. The span error occurred in the second example. Since '*both plasma and urine*' is content part, the FE should not include '*both*'. The other examples missed the CF-realising part. In the first example, '*a common approach*' is important to the introduction to the methodology. In the third example, detail number was extracted. It should be noted that the numbers sometimes constitute an FE because in some disciplines, there exist very fixed numbers, such as '*a p value less than 0.05 was considered significant*'. In the fourth example, the FE missed '*as proposed by*' to show the method was used in past work. In the last example, the controversy is represented by '*has been challenged*', which was not extracted.

The last example also shows that n-grams that contain the sentential root do not always convey the CF. It is true that the *that* clause conveys the CF *showing controversy within the field*, but the phrase in the main clause '*it should be noted that*' may have a different CF. This is a limitation when a sentence is regarded as a unit of a CF because a long sentence may have more than one CF. However, it is difficult to determine the length that constitutes the unit of a CF.

CF	Full sentence	Sentence without entities
Reference to tables or figures	From this table, we observe that the topics learned by our method are better in coherence than those learned from the baseline methods, which again demonstrates the effectiveness of our model.	from this * we observe that the topics learned by our * are better in * than those learned from the * which again demonstrates the * of our
Showing limitation or lack of past work	Although the cellular uptake efficiency could be improved by adjusting the size and the sequence of DNPs in the previous study, it has not been investigated whether the DNPs can also be used in the in vivo environment rich in nucleases.	although the * could be improved by adjusting the * and the * of * in the previous * it has not been * whether the * can also be used in the * rich in

Table 4: Examples of errors in named and scientific entity recognition. The sentences are cited from Xie, Yang, and Xing (2015); Kim et al. (2018).

CF	Sentence	FE
Showing brief introduction to the methodology	A common approach used to assign structure to language is to use a probabilistic grammar where each elementary rule or production is associated with a probability.	is to use a
Restatement of the results	For example, shared specific genomic aberrations were observed in both plasma and urine cfDNAs at loci of PTEN, TMPRSS2 and AR (Figure 1 and [CITATION]).	were observed in both
Description of the results	Rs679620 was also associated with increased OA risk in dominant (“TC-TT”, OR = 2.03, 95% CI: 1.03-4.01, P = 0.038) and overdominant model analyses (“TC”, OR = 2.04, 95% CI: 1.05-3.96, P = 0.033).	p 0038 and
Using methods used in past work	The smoothness value used for the AlphaSim calculation was based on the smoothness of the residual image of the statistical analysis as proposed by [CITATION] .	was based on the
Showing controversy within the field	However, it should be noted that the biological involvement of many of these targets in HBD-3 activities has been challenged in recent years [[CITATION]].	however it should be noted that the

Table 5: Examples of errors in n-gram extraction. The sentences are cited from Sarkar (1998); Xia et al. (2016); Guo et al. (2017); Vivas et al. (2019); Phan et al. (2016).

Table 6 shows the average number of FEs with 3/3 accuracy in each CF. It can be said that the difficulty in the FE extraction differs depending on the CFs. The CFs such as ‘describing interesting or surprising results’ and ‘unexpected outcome’ are often realised by an adverb or adjective, which is difficult to extract using the proposed method.

5.3 Error Analyses in Existing Methods

The existing FE extraction methods have different drawbacks. Table 7 lists the number of FEs extracted with the sentence-level methods after removing infrequent FEs occurring less than three times in the corpus. Compared to the proposed method, these methods extracted smaller numbers of FEs because most of these FEs rarely occur in the corpus. An example of sentence-level extraction is illustrated in Figure 3. The existing methods do not remove the non-formulaic words sufficiently here because the focus is only on a single word, and words such as ‘in’ or ‘results’ do not always constitute the FE.

The corpus-level methods are different in this regard. The numbers of extracted FEs are 23,847 (frequent n-gram) and 2,480,935 (Lattice FS). The frequent n-gram method extracts a smaller number of FEs because of the frequency thresholds. Further, it achieves a relatively good quality score, which is still lower than that of the proposed method (Table 2). The Lattice FS extracts too many FEs, which can deteriorate the quality of the FEs.

Original sentence	In order to avoid over fitting, PA with PCA was chosen for this study.
Frequency (1/50,000)	in order to avoid over fitting pa with * was chosen for this study
Frequency (1/100,000)	in order to avoid over fitting pa with pca was chosen for this study
LDA-based	in order to avoid over fitting * with * chosen for this study
Proposed	in order to avoid * was chosen for this

Figure 3: Example of FE extraction. The second step of the proposed method extracted two different n-grams. The original sentence is cited from An, Zhang, and Zhang (2018).

6 Conclusion

In this paper, we proposed a new sentence-level FE extraction method to realise CF-oriented analysis. We compared

CF	R.
Showing limitation or lack of past work	0.00
Comments on the findings	0.00
Showing explanation or definition of terms or notations	0.00
Unexpected outcome	0.00
Describing interesting or surprising results	0.00
Summary of the results	0.00
Comparison of the results	0.00
Showing the limitation of the research	0.00
Showing the characteristics of samples or data	0.00
Showing reasons why a method was adopted or rejected	0.00
Showing methodology used in past work	1.00
Suggestion of hypothesis	1.00
Showing the outline of the paper	1.00
Showing the aim of the paper	1.00
Suggestion of future work	1.00
Explanation for findings	1.00
Showing criteria for selection	1.00
Showing the main problem in the field	1.00

Table 6: CFs that the ratio of FEs with 3/3 accuracy is 0.00 or 1.00.

Method	FEs
Frequency-based (1/50,000)	13,722
Frequency-based (1/100,000)	12,840
LDA-based	18,033
Proposed	285,193

Table 7: Number of FEs that were extracted using the different methods and occurred at least three times in the dataset.

the proposed method to four existing methods, and our manual evaluations showed that the proposed method extracted CF-realising FEs better than these other methods. Although FE extraction has not been discussed in detail thus far in reported literature, we showed the existence of a more robust method than just extracting frequent n-grams, as adopted in the past studies. The FEs extracted with the proposed method are provided at our website[2] for utilisation in various tasks, such as information extraction and computer-based academic writing assistance.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19J12466 and 18H03297.

References

An, M.; Zhang, X.; and Zhang, X. 2018. Identifying the Validity and Reliability of a Self-Report Motivation Instrument for Health-Promoting Lifestyles Among Emerging Adults. *Frontiers in psychology* 9: 1222. ISSN 1664-1078. doi:10.3389/fpsyg.2018.01222.

Brooke, J.; Šnajder, J.; and Baldwin, T. 2017. Unsupervised Acquisition of Comprehensive Multiword Lexicons using Competition in an n-gram Lattice. *Transactions of the Association for Computational Linguistics* 5: 455–470. doi:10.1162/tacl_a_00073.

Cortes, V. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes* 12(1): 33–43. doi:10.1016/j.jeap.2012.11.002.

Guo, W.; Xu, P.; Jin, T.; Wang, J.; Fan, D.; Hao, Z.; Ji, Y.; Jing, S.; Han, C.; Du, J.; Jiang, D.; Wen, S.; ; and Wang, J. 2017. MMP-3 gene polymorphisms are associated with increased risk of osteoarthritis in Chinese men. *Oncotarget* 8(45): 79491–79497. doi:10.18632/oncotarget.18493.

Iwatsuki, K.; and Aizawa, A. 2018. Using Formulaic Expressions in Writing Assistance Systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2678–2689. Association for Computational Linguistics.

Kim, K.-R.; Röthlisberger, P.; Kang, S. J.; Nam, K.; Lee, S.; Hollenstein, M.; and Ahn, D.-R. 2018. Shaping Rolling Circle Amplification Products into DNA Nanoparticles by Incorporation of Modified Nucleotides and Their Application to In Vitro and In Vivo Delivery of a Photosensitizer. *Molecules* 23(7). ISSN 1420-3049. doi:10.3390/molecules23071833.

Liu, Y.; Wang, X.; Liu, M.; and Wang, X. 2016. Write-righter: An Academic Writing Assistant System. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 4373–4374. Association for the Advancement of Artificial Intelligence.

Mizumoto, A.; Hamatani, S.; and Imao, Y. 2017. Applying the Bundle–Move Connection Approach to the Development of an Online Writing Support Tool for Research Articles. *Language Learning* 67(4): 885–921. doi:10.1111/lang.12250.

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319–327. doi:10.18653/v1/W19-5034.

Phan, T. K.; Lay, F. T.; Poon, I. K.; Hinds, M. G.; Kvansakul, M.; and Hulett, M. D. 2016. Human β-defensin 3 contains an oncolytic motif that binds PI(4,5)P2 to mediate tumour cell permeabilisation. *Oncotarget* 7(2): 2054–2069. doi:10.18632/oncotarget.6520.

Sarkar, A. 1998. Conditions on Consistency of Probabilistic Tree Adjoining Grammars. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, 1164–1170. doi:10.3115/980691.980759.

Simpson-Vlach, R.; and Ellis, N. C. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4): 487–512. ISSN 0142-6001. doi:10.1093/applin/amp058.

Vivas, A. B.; Paraskevopoulos, E.; Castillo, A.; and Fuentes, L. J. 2019. Neurophysiological Activations of Predictive and Non-predictive Exogenous Cues: A Cue-Elicited EEG Study on the Generation of Inhibition of Return. *Frontiers in Psychology* 10: 227. ISSN 1664-1078. doi:10.3389/fpsyg.2019.00227.

Xia, Y.; Huang, C.-C.; Dittmar, R.; Du, M.; Wang, Y.; Liu, H.; Shenoy, N.; Wang, L.; ; and Kohli, M. 2016. Copy number variations in urine cell free DNA as biomarkers in advanced prostate cancer. *Oncotarget* 7(24): 35818–35831. doi:10.18632/oncotarget.9027.

Xie, P.; Yang, D.; and Xing, E. 2015. Incorporating Word Correlation Knowledge into Topic Modeling. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 725–734. doi:10.3115/v1/N15-1074.

[2] https://github.com/Alab-NII/CF-Labelled-FE-Database