Using Formulaic Expressions in Writing Assistance Systems

Kenichi Iwatsuki[†]
[†]The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, Japan
iwatsuki@nii.ac.jp

Akiko Aizawa[‡†]
[‡]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, Japan
aizawa@nii.ac.jp

Abstract

Formulaic expressions (FEs) used in scholarly papers, such as '*there has been little discussion about*', are helpful for non-native English speakers. However, it is time-consuming for users to manually search for an appropriate expression every time they want to consult FE dictionaries. For this reason, we tackle the task of semantic searches of FE dictionaries. At the start of our research, we identified two salient difficulties in this task. First, the paucity of example sentences in existing FE dictionaries results in a shortage of context information, which is necessary for acquiring semantic representation of FEs. Second, while a semantic category label is assigned to each FE in many FE dictionaries, it is difficult to predict the labels from user input, forcing users to manually designate the semantic category when searching. To address these difficulties, we propose a new framework for semantic searches of FEs and propose a new method to leverage both existing dictionaries and domain sentence corpora. Further, we expand an existing FE dictionary to consider building a more comprehensive and domain-specific FE dictionary and to verify the effectiveness of our method.

Title and Abstract in Japanese

執筆支援システムにおける定型表現の利用

論文中に用いられる定型表現（formulaic expressions）とは，例えば'there has been little discussion about'などであり，非英語母語話者の英語論文執筆に有益である。しかしながら，定型表現辞書をその都度参照して適切な表現を探すのは容易ではない。そこで本研究では，計算機による定型表現辞書の意味検索に取り組む。まず，本タスクにおける課題として以下の2つを挙げる。既存の定型表現辞書に収録されている例文が少なく意味表現の獲得に必要な文脈情報が不足していること，多くの定型表現辞書では意味カテゴリごとに定型表現が分類されているが，ユーザの入力から意味カテゴリを予測することが困難であり，ユーザに意味カテゴリを明示する手間が発生することである。 これらの課題を解決するため，本研究では定型表現の意味検索の新たな枠組みを提案し，既存の定型表現辞書を分野毎の英文コーパスと対応づけて利用するための手法を提案する。 さらに，既存の定型表現辞書を拡張し，より網羅性が高く，かつ分野特化型の定型表現辞書の構築を検討し，有効性を検証する。

1 Introduction

Non-native English speakers writing scholarly papers often use the same expressions repeatedly or are not confident in the correctness of their usage of certain wording. Existing computer-based writing assistance systems do not always help them find better expressions than those they already know because such systems search a corpus by using simple pattern-matching based on input keywords (Chang et al., 2015; Chang and Chang, 2015; Jeong et al., 2014; Liu et al., 2016; Mizumoto et al., 2017). For instance, when a user wants to write about previous work and find expressions other than '*little research was done* *on*', the only one the user might know, a keyword-based search does not present expressions such as '*there has been little discussion about*' to the user.

A reasonable solution is using dictionaries of formulaic expressions (FEs) (Ellis et al., 2008; Conklin and Schmitt, 2008). FEs used in scholarly papers are helpful for non-native English speakers writing papers in English (Peters and Pauwels, 2015). FEs have a communicative function and thus in FE dictionaries, they are classified into categories on the basis of their functions. Consequently, if a user wants to write about the fact that there is little research on a topic, the user goes to the category '*to show lack of existing research*' and finds expressions such as '*there has been little discussion about*', even if the user has not ever encountered the entire expression.

Despite the usefulness of such FE dictionaries, it is time-consuming for users to choose a category and manually find candidate expressions every time they consult the dictionary. Therefore, the aim of our research is to enable semantic searches of FE dictionaries by computers with incomplete user input. In this paper, we formulate this problem as a prediction task for categories in FE dictionaries. Note that user input can be an incomplete sentence that does not contain a FE because we suppose users who do not think of FEs or who just input some words they have in mind.

We started by identifying two salient difficulties. The first is that very few example sentences are included in FE dictionaries, resulting in the lack of context information necessary for acquiring semantic representation of FEs. The second is that writing assistance systems must automatically predict a semantic category from user input so that users no longer need to designate the category.

In this paper, we propose a method to create a new writing assistance system and verify if it works. First, we extracted example sentences from corpora for a given FE. Second, we checked if having sufficient example sentences improved the prediction of a category from user input that would be an incomplete sentence or phrase. Finally, we classified FEs into categories using context information including full sentences and section labels in addition to the FEs themselves. In our experiment, we used two corpora from different disciplines, one from ACL Anthology (computational linguistics) and the other from PubMed (life science & medicine). For an existing FE dictionary, we used Academic Phrasebank (Morley, 2018).

FEs recorded in the dictionaries are not always suitable for writing assistance systems because many of the expressions do not occur in an actual corpus and differ depending on disciplines. Thus, we aim to build a more comprehensive and more domain-specific FE dictionary by expanding an existing FE dictionary. In this study, we formulate the extraction of FEs from corpora as a sequential labelling problem, which is solved by learning *formulaicity* with an existing dictionary.

The contributions of this paper are as follows. We present a new framework for semantic searches of FEs that can be used when writing scientific papers. We propose leveraging both existing dictionaries and domain-sentence corpora and show that using context information extracted from actual corpora improves the category-prediction accuracy, compared to using the context information originally recorded in the existing FE dictionary. We reformulate FE extraction as a sequential-labelling problem and show that the quality of the FEs extracted with our method is higher than that of those with previous methods.

2 Related Work

2.1 Writing Assistance Systems

Existing writing assistance systems are classified into three types. First, the most direct approach for computer-based writing assistance is that in which user-input sentences are used to retrieve example sentences. Search results are shown with concordances (Wu et al., 2006) or dependency structures (Kato et al., 2006).

Another approach is similar to an input method in which users can input non-alphabetical languages. FLOW (Chen et al., 2012) suggests an English translation from words written in another language. WINGS (Dai et al., 2014) suggests full Chinese sentences and words from pinyin. Full sentences are suggested on the basis of searches for sentences that contain words that are the same as or similar to the input.

The third approach is combined with an authoring system. With this approach, candidate English expressions that follow user input are listed; then the users can choose one of them (Jeong et al., 2014; Chang et al., 2015; Yen et al., 2015; Chang and Chang, 2015; Liu et al., 2016; Mizumoto et al., 2017). Some systems allow users to specify the categories of FEs. Such categories include the Introduction, Method, Results and Discussion (IMRaD) structure (Jeong et al., 2014), argumentative zone (Teufel, 1999; Chang et al., 2015) and move-step structure (Swales, 1990; Mizumoto et al., 2017). Note that users must designate which category to use.

Overall, existing writing assistance systems adopt keyword-based searches, and thus the number of suggested expressions is limited to ones that contain user-input keywords.

2.2 Definitions of Formulaic Expressions

A survey of definitions of FEs shows that there are three ways of defining them (Durrant and Mathews-Aydınlı, 2011). The first is a 'phraseological' approach. Using this approach, *formulaicity* is definable by non-compositionality of word sequences. However, this definition is not for FEs but for idioms because the semantics of FEs are often compositional. For example, '*have been explored by many researchers*' has a compositional meaning but it is nonetheless a FE. The second is a 'frequency-based' approach. In this approach, frequently co-occurring word sequences are considered FEs. However, noise such as '*is one of the*' cannot be removed. Also, FEs do not always occur frequently. The third one is a 'psychological' approach, which defines FEs as word sequences that are processed and remembered as a whole in the human brain. This seems to be a successful definition of formulaicity, but computers cannot process word sequences using this approach.

Several analyses of FEs exist. Biber et al. (2004) analysed the usage of lexical bundles (continuous word sequences) in an academic context. They defined lexical bundles as 'the most frequent recurring lexical sequences in a register'. Their results showed that lexical bundles are not always syntactically structured. In fact, they often contain some fragments such as '*is based on the*', '*I don't know if*' and '*a little bit of*'.

Along with lexical bundles, Gray and Biber (2013) specifically examined phrase frames: discontinuous word sequences with a slot ' * 'that is filled by any word. The number of lexical bundles used in corpora is larger than that of phrase frames, but examining particularly those occurring in at least five texts, phrase frames are more numerous than lexical bundles. They classified phrase frames into three types: verb-based frames, frames with other content words and function word frames.

Although there are several definitions of FEs, many dictionaries of FEs has been published in which FEs are collected on the basis of intuitive definitions.

2.3 Extraction of Formulaic Expressions

Simpson-Vlach and Ellis (2010) extracted word sequences, three to five words in length, from corpora. Then, with frequency and mutual information, the extracted word sequences were ranked. Their results show that highly frequent word sequences alone cannot be regarded as FEs because they have no distinctive function or meaning. Further, useful expressions are not always highly frequent. In fact, word sequences with high mutual information are rare in corpora because many are subject-specific.

Vincent (2013) decomposed a candidate phrase into the phrasal core and its collocates. The phrasal core is a continuous or discontinuous word sequence occurring with high frequency. Candidate phrases including the core were first identified in a corpus. Then the collocates were sought.

Brooke et al. (2015) used the technique for multi-word expression extraction (Brooke et al., 2014) to find FEs. They split a sentence into parts with a lexical predictability ratio. They pointed out that evaluating newly acquired FEs is difficult because there is no answer dataset.

Aside from studies of FE extraction, several studies have addressed phrase extraction in particular for information extraction or text mining (Zhong et al., 2012; Zhang et al., 2013; Liu et al., 2015). However, these studies specifically addressed characteristic phrases that were informative. Therefore, FEs such as '*in this paper*' were not considered target expressions.

3 Semantic FE Search for Writing Assistance System

3.1 Proposed Searching System

Most existing writing assistance systems seek candidate expressions using keyword-based searches, which causes the problem wherein expressions very different from the input are not presented to users. Consequently, we first consider a semantic searching system that searches for FEs. Then, we propose a scheme for a dictionary of FEs.

The proposed searching system is based on the way people manually use FE dictionaries. Specifically, a user chooses a category that expresses the user's intent. Then, the user picks one of the FEs belonging to that category. Consequently, the proposed searching process consists of two steps. First, the system presumes a user's intention on the basis of written words, which may be an incomplete phrase or sentence. Second, the system searches a dictionary for candidate expressions in the category corresponding to the presumed intention. To choose the correct category, the system must use information related to context that is derived from what a user is writing. Therefore, a resource must include information related to context in which an expression is used. For example, the expression '*further research should be undertaken to investigate*' is not suggested by a keyword-based search with the keywords '*it is future work to investigate*' but by category-based search with the category '*about future work*'.

For this system, the proposed dictionary is formalised as follows. The dictionary consists of several categories, and each category has multiple FEs. Each FE has one or more example sentences. Figure 1 shows an image of the whole dictionary.

In the following subsections we discuss categories and FEs.

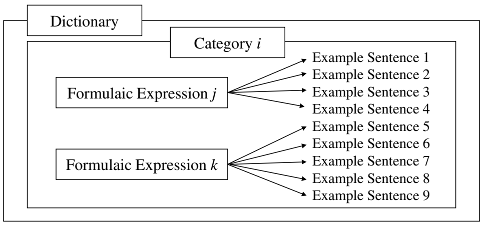

Figure 1: In the dictionary, categories have several FEs, each of which has corresponding example sentences.

3.2 Categorisation Systems

The question remains of what structures of categories are suitable for writing assistance. Presuming that FEs are classified according to their functions, writing assistance systems must anticipate what users want to write from the surrounding context. However, specifically examining scholarly papers shows that the structure of scientific documents is fixed to some degree. Consequently, as long as users write along this fixed structure, there is a good chance that the system can anticipate what the user wants to write if the categories correspond to the paper structure. Therefore, on the basis of the logical structure, categories correspond to both the functions of FEs and the user's intention.

A section-based structure is the simplest, with the Introduction, Method, Results and Discussion (IMRaD) structure adopted for many papers. However, the sections in a paper are so few that too many FEs belong to one category to choose an appropriate one easily. As described herein, we particularly examine move-step structures proposed by Swales (1990). According to Swales' analyses, a paper is composed of several sections, which include moves, each of which has steps (Table 1).

Several analyses have assessed move-step structures in scientific papers. In *Introduction* sections, the Create-A-Research-Space model was found to be adopted in many papers (Swales, 1990; Swales, 2004). Some research has specifically addressed move structures in all sections (Cotos et al., 2015) or *Abstracts* (Lorés, 2004). Other studies have emphasised examining the transition between moves (Ozturk, 2007) and differences in usage across disciplines (Peacock, 2002; Maswana et al., 2015).

Section: Introduction		
Move 1: Establishing a territory	Move 2: Establishing a niche	Move 3: Occupying the niche
Step 1: Claiming centrality	Step 1A: Counter-claiming	Step 1A: Outlining purposes
Step 2: Making topic generalization	Step 1B: Indicating a cap	Step 1B: Announcing present research
Step 3: Reviewing items of previous research	Step 1C: Question-raising	Step 2: Announcing principal findings
	Step 1D: Continuing a tradition	Step 3: Indicating RA structure

Table 1: A move-step structure for the introduction section of a research article (RA) called the Create-A-Research-Space model proposed by Swales (1990).

In our research, we use categories based on move-step structures for writing assistance systems. Specifically, we use the categorisation system that is adopted in Academic Phrasebank made by Morley (2018) because the categorisation of this resource is similar to move-step structures. In Academic Phrasebank there are six sections: Introducing Work, Referring to Sources, Describing Methods, Reporting Results, Discussing Findings and Writing Conclusions and 77 categories such as *Establishing the importance of the topic for the discipline* and *Giving reasons why a method was adopted or rejected*, which roughly correspond to steps. This is suitable for scholarly articles.

3.3 Formulaic Expressions

Considering previous work on the definitions of FEs and our semantic search model, we decide our target FEs as follows. FEs must be expressions that are helpful in writing a paper. Word sequences with an unclear function are excluded. For example, '*is the number of*' and '*the word in the*' do not have clear functions. Other expressions that should be excluded are classified into two types: overly general and overly specific expressions. The former type consists of phrases such as '*on the other hand*' and '*we would like to*'. These might appear to be useful, but they cannot be assigned one category label because '*on the other hand*' can be used in many sections of a paper. Therefore, we chose to exclude them from this research. '*Natural language processing*' or '*in the training dataset*' are examples of the latter type. Overall, in our research, we defined FEs as word sequences whose functions correspond to one category.

4 Methods

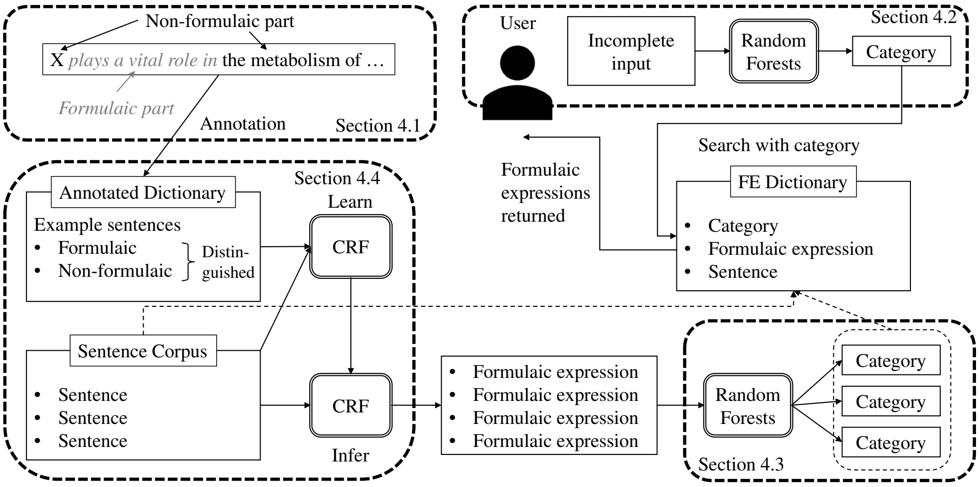

Figure 2: In Section 4.1, annotation is conducted to distinguish the formulaic from the non-formulaic part in each expression recorded in the existing dictionary to build the annotated dictionary. In Section 4.2, incomplete user input is assigned a category label. In Section 4.3, FEs and sentences are classified into categories. In section 4.4, FEs are extracted from a sentence corpus by learning formulaicity from the annotated dictionary.

4.1 Dictionary Annotation

We used two kinds of language resources: an existing dictionary of FEs and a sentence corpus extracted from a group of scientific papers. We first built an annotated dictionary from the existing dictionary.

Many expressions are recorded in the existing dictionary, but the formulaic and non-formulaic parts are not distinguished. For example, '*X plays a vital role in the metabolism of ...*' contains non-formulaic parts such as 'X', 'metabolism' and '...'. This is most problematic when using the existing dictionary because we cannot search corpora for sentences that contain FEs, using only formulaic parts. Therefore, for formulaicity to be learned from the dictionary, we first annotated the formulaic parts in the dictionary. Each word of an expression was manually labelled as either formulaic or non-formulaic (upper-left in Figure 2). Then, non-formulaic parts were removed. After annotation, we extracted sentences containing the annotated FEs from a corpus. Finally, in this annotated dictionary, each categorised FE has some example sentences derived from an actual corpus (Figure 3). All words are lemmatised to avoid inflections. Annotation guidelines are available on our GitHub repository[1].

Category: Stating the purpose of research
Formulaic expression: the objective of this research be to
Example sentences from ACL Anthology:
one of *the objectives of this research is to* make online documents more understandable by paraphrasing unknown …
Thus *the objective of this research is to* experiment with these techniques for Sinhala-Tamil, and identify the best …

Figure 3: In the annotated dictionary, each FE is classified into a category and has several example sentences extracted from a corpus[3].

As described in Section 3.2, we used Academic Phrasebank (Morley, 2018) as a dictionary because its categories are based on move-step structures.

4.2 Prediction of Categories from Incomplete Sentences

In the proposed searching system described in section 3.1, categories are automatically predicted from user input. This is formalised as a classification task from user input into categories.

The prediction is conducted with random forest[4](upper-right in Figure 2).

Each user-input is represented as a vector, which is an average of skip-gram vectors (Mikolov et al., 2013) of each word in the input. The vector representations are learned with a sentence corpus.

The annotated corpus is so divided into a training dataset and a test dataset that sentences in one category in one dataset are almost as many as those in the other dataset.

User input can be an incomplete sentence. Therefore, inputs of two types are prepared as presumed user inputs. One is a sentence not containing a FE, based on the idea that a user wants to find a FE with content words (WithoutFE). The other is a sentence with half or two thirds of the composing words removed. Specifically, we simply select words from every two or three words in a sentence. This simulates a situation where a user comes up with some words but does not compose a sentence (PickedWords). This vector can include part of a FE because users may know some of the words composing the FE.

4.3 Classification of FEs into Categories

In the proposed FE dictionary, FEs are classified into categories. In the previous section classified user input into categories; this section classifies FEs into categories. Therefore, the same methodology as the previous section can be applied to this task, but the input features are different.

We used random forests as a classifier and the input was a vector representation of a FE. The output was a category label (lower-right in Figure 2).

[1] `https://github.com/Alab-NII/FE/blob/master/annotation_guidelines.md`

[3] The example sentences were extracted from two papers: Higashinaka, R., & Nagao, K. 2002. Interactive paraphrasing based on linguistic annotation. In *COLING 2012*. and Hameed, R. A., et al. 2016. Automatic Creation of a Sentence Aligned Sinhala-Tamil Parallel Corpus. In *WSSANLP 2016*.

[4] we use an R package (`https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`).

We used a sentence and a section label in which a FE is used as context information. We tested five kinds of vectors, as described below. Each vector is an average of the vectors of each composing word.

Sentence Vector (Sen) This vector is produced from every word of a sentence that includes the FE.

Sentence Vector without Functional Word Effects (Sen - FW) This vector was presented by Arora et al. (2017). To decrease the effect of functional words, each word embedding is weighed by the coefficient $\frac{a}{a+p}$ where a denotes a parameter set as 0.001 in our experiment and p is word frequency. Afterwards, singular value decomposition is applied.

Sentence Vector without Content Word Effects (Sen - CW) To decrease the content word effects, words with low frequency are removed. When calculating the average vector of each word embedding, words that occur in the sentence corpus less than once per million words are ignored.

Sentence Vector with Section Information (Sen + Sec) A simple one-hot vector which shows that the title of a section in which the FE is used contains *introduction* or *background*, *related work*, *method* or *approach*, *experiment* or *evaluation*, *result*, *discussion* or *conclusion* or*future*, is concatenated to the sentence vector. Some FEs appear in multiple sections, so we assigned the most popular section label in each FE to sentences.

Formulaic Expression Vector (FE) This vector is a baseline made solely from words that compose the FE in the annotated dictionary, which means that context information is excluded.

4.4 Acquisition of new FEs from Corpora

To exploit the intuitive definition of formulaicity in Academic Phrasebank, we implemented a supervised learning method to extract FEs. The FEs recorded in Academic Phrasebank and occurring in real corpora are few in number, so we first paraphrased all nouns, adjectives and adverbs in the annotated dictionary to expand the dataset, using PPDB2.0 (S size) (Pavlick et al., 2015). Subsequently, we attached BIO tags to each sentence in the annotated dictionary. Here, the first word of a FE is assigned a B label. Other words of a FE are tagged with an I label. The rest of the words in the sentence are labelled O. Formulaicity is learned using these sentences. We used conditional random fields (CRFs)[5] to learn and extract FEs. Features were words, part-of-speech tags and word frequency, which was discretised. Then, using the learned model, each word of every sentence in a sentence corpus was given a BIO tag (lower-left in Figure 2). This formulation facilitates the extraction of both continuous and discontinuous FEs.

5 Evaluation

5.1 Datasets

We built two kinds of datasets: an annotated dictionary and a sentence corpus. To build the annotated dictionary, we annotated Academic Phrasebank. Three annotators annotated FEs in five categories to check inter-annotator agreement. The value was 0.699 (Cohen's kappa coefficient). Then, the rest of the FEs were annotated by one annotator. Academic Phrasebank as an existing dictionary had 77 categories, but some categories in which there are fewer than two example sentences extracted from the sentence corpus or in which no FE was found after the annotation were removed. Consequently, 39 categories for ACL Anthology and 45 categories for PubMed were adopted.

We compiled a sentence corpus for which sentences were extracted from papers in ACL Anthology and PubMed. Our corpus consisted of 2,629,115 sentences from ACL Anthology and 2,894,721 sentences from PubMed. Table 2 shows the number of FEs and sentences after the paraphrasing. By paraphrasing FEs and sentences were increased. However, while the number of FEs in the annotated dictionary was originally 854, most of them did not occur in the corpora so that the number was reduced to 225 and 334, respectively.

[5] We use CRF++ (`https://taku910.github.io/crfpp/`).

ACL / PubMed	Before paraphrasing	After paraphrasing
Number of formulaic expressions	225 / 334	352 / 540
Number of example sentences	40,510 / 66,193	49,118 / 89,734

Table 2: Sentence corpora characteristics show that the number of FEs and sentences increased by paraphrasing.

5.2 Prediction of Categories from Incomplete Sentences

To show that sufficient example sentences are necessary for the writing assistance system to predict a category, we classified the presumed user input into a category with a comparison to the original Academic Phrasebank and the annotated dictionary.

Table 3 presents the accuracy, macro precision, macro recall and macro F-measure of the prediction. These results indicate that example sentences recorded in Academic Phrasebank are insufficient for the writing assistance system to predict a category from user input. Consequently, the dictionary requires example sentences extracted from an actual corpus depending on the writer domain.

Dataset	ACL Anthology				PubMed			
	Acc	P	R	F	Acc	P	R	F
WithoutFE w/AP	1.64%	3.52%	6.44%	0.64%	0.69%	2.61%	4.45%	0.52%
PickedWords w/AP	11.2%	6.04%	5.52%	3.14%	3.32%	4.91%	3.72%	2.21%
WithoutFE w/SC	55.9%	77.1%	20.6%	27.3%	47.8%	55.1%	17.8%	20.0%
PickedWords w/SC	81.7%	90.0%	43.2%	54.5%	72.1%	77.9%	37.6%	45.8%

Table 3: Accuracy (Acc), average precision (P), average recall (R) and average F-measure (F) of the classification are shown. The values in upper two rows are the results of the classification in which the training dataset was made from only example sentences in Academic Phrasebank (AP), while those in the lower two rows are the result of the proposed method trained with many more example sentences in a sentence corpus (SC).

5.3 Classification of FEs

The number of FEs was small, so we did a leave-one-out cross validation with FE vectors. The other types of vectors described in Section 4.3 were made from sentences. Eventually, the annotated corpus was divided into training and test datasets in the same way as described in Section 4.2. We also used Academic Phrasebank sentences as a training data for comparison.

The results are presented in Table 4. The results show that the classification of FEs is improved when context information such as example sentences and section information is used. Additionally, removing the effect of highly frequent word adversely affects the classification accuracy, probably because FEs usually consist of frequent words. The results demonstrate that sentences recorded in Academic Phrasebank are insufficient for the writing assistance system to classify FEs into categories.

5.4 Acquisition of new FEs

Using the learned model with CRFs, we extracted FEs from the corpora. After extraction, word sequences that occurred less than twice in the corpus, with a length of less than four words or in which the same words (excluding prepositions) were used twice or more were removed. From ACL Anthology we obtained 2,086 FEs (including 481 discontinuous FEs) with 117,889 example sentences. From PubMed we acquired 1,884 FEs (including 172 discontinuous FEs) with 184,311 example sentences. The extracted FEs are available on our GitHub repository[6].

We manually evaluated each extracted FE, checking if it met the definition we made. We picked FEs in the following way. First, we calculated the smallest edit distances between each FE and all expression

[6] `https://github.com/Alab-NII/FE/`

	ACL Anthology				PubMed			
Input vector	Acc	P	R	F	Acc	P	R	F
Sen	71.7%	79.0%	28.4%	35.7%	61.1%	51.1%	21.0%	23.3%
Sen - FW	65.1%	77.1%	25.3%	32.7%	54.2%	48.3%	18.6%	20.7%
Sen - CW	72.0%	78.3%	27.7%	35.0%	61.5%	51.7%	21.2%	23.4%
Sen + Sec	77.5%	78.7%	33.3%	39.8%	79.6%	89.4%	60.2%	70.2%
Sen - CW + Sec	77.9%	79.8%	34.6%	41.3%	79.8%	89.2%	60.7%	70.6%
FE	45.2%	26.3%	26.4%	24.1%	48.5%	32.9%	33.8%	32.0%
Majority (baseline)	33.3%	0.85%	2.56%	1.28%	12.3%	0.27%	2.22%	0.49%
Sen - CW + Sec (no SC)	8.95%	7.11%	9.56%	2.98%	4.72%	8.34%	5.81%	2.09%

Table 4: The upper five rows are the results of the classification with the proposed vectors. For comparison, three types of experiments were also conducted: one in which the vectors were made only from formulaic expressions (FE), another is in which they were made only from voting for majority class and the third in which Sen - CW + Sec vectors were learned only with example sentences in Academic Phrasebank, without using a sentence corpus (no SC).

	ACL Anthology		PubMed	
	Cont. FE	Discont. FE	Cont. FE	Discont. FE
Our method	54% (60/111)	32% (8/25)	67% (62/93)	26% (10/37)
High frequency	14% (15/110)	-	14% (14/102)	-
High MI	15% (16/106)	-	12% (12/100)	-

Table 5: Our method outperforms mere extraction of word sequences with high frequency or mutual information (MI).

in the annotated dictionary. Then, we randomly chose FEs in each class of the distance. This is based on the idea that FEs similar to those originally recorded in Academic Phrasebank are more likely to meet the criteria than FEs quite different from them. Three annotators conducted the evaluation with small samples to calculate inter-annotator agreement: 0.800. About 100 FEs were evaluated by one annotator. For comparison, we extracted word sequences with high frequency or high mutual information, which was adopted by Simpson-Vlach and Ellis (2010).

Table 5 presents the ratio of good FEs in each result. Comparing with previous methods, the proposed method extracted continuous FEs more successfully. Discontinuous FEs are more difficult to extract correctly. Some examples are shown in Table 6.

6 Discussion

6.1 Error Analysis of Extracted FEs

We analysed the errors in the extracted continuous FEs. We first divided errors into two categories: spanning and semantics errors. The former occur when one word is unnecessary or when the words are insufficient to make a good FE. Consequently, errors of four types occur, one in which a word to the left or right is unnecessary or missing. The remaining errors are classified into three types: too general, too specific and nonsensical. Overly general and overly specific FEs were explained in Section 3.3. Nonsensical expressions are word sequences that are unlikely to be helpful in writing.

The distribution of errors is presented in Figure 4. In any case nonsensical errors are the most frequently occurring, but the remaining errors differ across datasets and methodologies. Examining the errors made in ACL Anthology using our method, unnecessary words (right) and insufficient words (left) stand out. The former error occurs mainly when 'if' is included in a FE, such as *there be evidence to suggest that if*. The latter error includes phrases such as *be important to note that*. Most missing words are 'it'. However, 'if' is sometimes necessary and 'it' is sometimes unnecessary. For example, when 'if' can be replaced with 'whether' in expressions such as *check if*, appending 'if' to a FE is very helpful.

	ACL Anthology	PubMed
Our method	have(has) shown significant improvement(s) over	previous studies have demonstrated
	is(are) a key component in	the aim of this study is to investigate
	it be possible to * to determine whether	it is possible that * may have influence
	to that of Figure	*the number of investigated*
	has been observed that the thematic role	*to measure * suffer from*
High freq.	we can see that	it is important to
	in this paper we propose	play an important role
	state of the art	*(Figure 1)*
	shown in Figure 2	*et al 2010)*
	the words in the	*$p < 0.05$)*
High MI	this paper is organized as follows	it is important to
	to the best of our knowledge	play an important role
	statistical machine translation (SMT)	*of this study is to*
	on the other hand	*et al 2011)*
	a large number of	*($p < 0.05$)*

Table 6: Examples of extracted FEs. Expressions that are not regarded as good FEs are written in italics.

Additionally, ‘it’ as a formal subjective should appear along with a real subject, such as a to-infinitive or a that-clause. However, in the PubMed dataset, insufficient words (left) stand out. The words missing the most are ‘it’ and ‘there’. For example, ‘it’ should be appended to *be clear from the figure that* and *be important to investigate if* and ‘there’ to *be significant difference between* and *be some limitation to this study*.

Specifically examining FEs extracted using the existing methods, it is apparent that there are many overly general expressions. These expressions mainly include prepositional and idiomatic phrases such as *on the basis of*, *at the time of*, *on the other hand*, *be more likely to* and *for the purpose of*. A different categorisation system is necessary to make the use of these expressions.

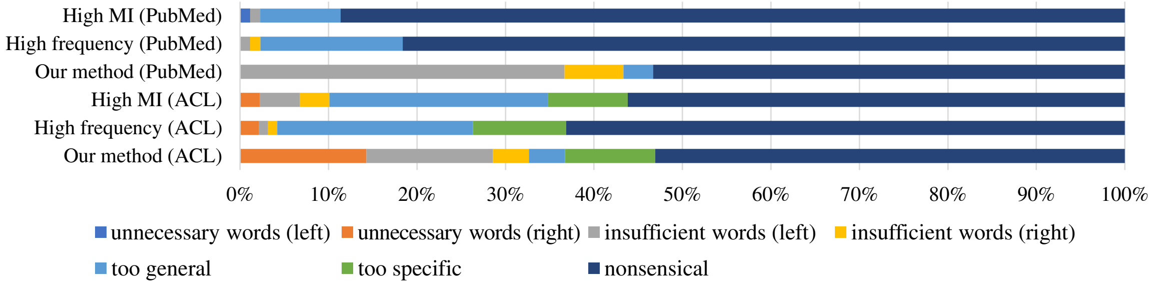

Figure 4: Distribution of error types among results.

6.2 Difference between Domains

For this work, we used two corpora, ACL Anthology and PubMed. Here we discuss the differences in the extracted FEs between the two corpora. First, there were 339 FEs occurring in both corpora (16% of ACL Anthology; 18% of PubMed).

Domain-specific technical terms, such as *natural language processing* or *reactive oxygen species*, are unlikely to be extracted using the proposed method. However, the usage of FEs is shown to differ across domains, which implies that the expressions should be re-ranked according to the users’ discipline when candidate expressions are presented to users of the writing assistance system. In addition, it is interesting to note that section information seems more critical in PubMed’s classification than in that of ACL.

7 Conclusion

We proposed a new framework for semantic searches of FEs with incomplete user input. We also proposed a method to classify FEs into categories and showed that context information improves the accuracy of the classification. Further, we reformulated FE extraction as a sequential labelling problem and found that this method works well to build a domain-specific FE dictionary.

On the basis of our approach, further research will be conducted to improve the classification and extraction. For classifying FEs, vector representations of a sentence should probably be improved focusing on FE. For acquiring new FEs, an advanced machine learning algorithm should be used to reduce nonsensical FEs and syntactic information would be useful to alleviate the spanning problem.

Acknowledgements

This research was supported by JSPS KAKENHI 18H03297.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371–405.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 753–761.

Julian Brooke, Adam Hammond, David Jacob, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2015. Building a lexicon of formulaic language for language learners. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 96–104.

Jim Chang and Jason Chang. 2015. Writeahead2: Mining lexical grammar patterns for assisted writing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 106–110.

Jim Chang, Hsiang-Ling Hsu, Joanne Boisson, Hao-Chun Peng, Yu-Hsuan Wu, and Jason S. Chang. 2015. Learning sentential patterns of various rhetoric moves for assisted academic writing. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 37–45.

MeiHua Chen, ShihTing Huang, HungTing Hsieh, TingHui Kao, and Jason S. Chang. 2012. Flow: A first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 System Demonstrations*, pages 157–162.

Kathy Conklin and Norbert Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1):72–89.

Elena Cotos, Sarah Huffman, and Stephanie Link. 2015. Furthering and applying move/step constructs: Technology-driven marshalling of swalesian genre theory for eap pedagogy. *Journal of English for Academic Purposes*, 19(Supplement C):52–72. 25 Years of “Genre Analysis”.

Xianjun Dai, Yuanchao Liu, Xiaolong Wang, and Bingquan Liu. 2014. Wings:writing with intelligent guidance and suggestions. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30.

Philip Durrant and Julie Mathews-Aydınlı. 2011. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1):58–72.

Nick C. Ellis, Rita Simpson-vlach, and Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3):375–396.

Bethany Gray and Douglas Biber. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1):109–136.

Senator Jeong, Sejin Nam, and Hyun-Young Park. 2014. An ontology-based biomedical research paper authoring support tool. *Science Editing*, 1(1):37–42.

Y. Kato, S. Matsubara, and Y. Inagaki. 2006. A corpus search system utilizing lexical dependency structure. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744.

Yuanchao Liu, Xin Wang, Ming Liu, and Xiaolong Wang. 2016. Write-righter: An academic writing assistant system. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 4373–4374.

Rosa Lorés. 2004. On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes*, 23(3):280–302.

Sayako Maswana, Toshiyuki Kanamaru, and Akira Tajino. 2015. Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand*, 2:1–11.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Atsushi Mizumoto, Sawako Hamatani, and Yasuhiro Imao. 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.

John Morley. 2018. Academic phrasebank. `http://www.phrasebank.manchester.ac.uk/`.

Ismet Ozturk. 2007. The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes*, 26(1):25–38.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430.

Matthew Peacock. 2002. Communicative moves in the discussion section of research articles. *System*, 30(4):479–497.

Elke Peters and Paul Pauwels. 2015. Learning academic formulaic sequences. *Journal of English for Academic Purposes*, 20:28–39.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4):487–512.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

John Swales. 2004. *Research genres: Explorations and applications*. Cambridge University Press.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Benet Vincent. 2013. Investigating academic phraseology through combinations of very frequent words: A methodological exploration. *Journal of English for Academic Purposes*, 12(1):44–56.

Jien-Chen Wu, Yu-Chia Chang, Hsien-Chin Liou, and Jason S. Chang. 2006. Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 41–44.

Tzu-Hsi Yen, Jian-Cheng Wu, Jim Chang, Joanne Boisson, and Jason Chang. 2015. Writeahead: Mining grammar patterns in corpora for assisted writing. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 139–144.

Bin Zhang, Alex Marin, Brian Hutchinson, and Mari Ostendorf. 2013. Learning phrase patterns for text classification. *IEEE transactions on audio, speech, and language processing*, 21(6):1180–1189.

Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. 2012. Effective pattern discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*, 24(1):30–44.