An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers

Kenichi Iwatsuki, Florian Boudin, Akiko Aizawa

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1712–1720
Marseille, 11–16 May 2020

An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers

Kenichi Iwatsuki[1], **Florian Boudin**[2], **Akiko Aizawa**[31]
[1]The University of Tokyo, 7-3-1, Hongo, Bunkyo, 113-8656 Tokyo, Japan
[2]Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France
[3]National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda, 101-8430 Tokyo, Japan
iwatsuki@nii.ac.jp, florian.boudin@univ-nantes.fr, aizawa@nii.ac.jp

Abstract

Formulaic expressions, such as '*in this paper we propose*', are used by authors of scholarly papers to perform communicative functions; the communicative function of the present example is '*stating the aim of the paper*'. Collecting such expressions and pairing them with their communicative functions would be highly valuable for various tasks, particularly for writing assistance. However, such collection and paring in a principled and automated manner would require high-quality annotated data, which are not available. In this study, we address this shortcoming by creating a manually annotated dataset for detecting communicative functions in sentences. Starting from a seed list of labelled formulaic expressions, we retrieved new sentences from scholarly papers in the ACL Anthology and asked multiple human evaluators to label communicative functions. To show the usefulness of our dataset, we conducted a series of experiments that determined to what extent sentence representations acquired by recent models, such as word2vec and BERT, can be employed to detect communicative functions in sentences.

Keywords: multi-word expression, formulaic expression, communicative function, rhetorical structure, sentence representation

1. Introduction

Formulaic expressions are defined as 'multi-word expressions that speakers remember and process as wholes, rather than constructing them "online" with each use' (Durrant and Mathews-Aydınlı, 2011). They are frequently used in scientific papers because they convey specific communicative functions in the rhetorical structure of papers, which indicate the author's purpose or intention (Swales, 1990; Swales, 2004). For example, the formulaic expression '*in this paper we propose*' has the communicative function '*stating the aim of the paper*'. There is, however, no consensus as to what constitutes the minimal text span that realises a communicative function. For example, to convey the function '*describing the limitations of current research*', some may regard '*beyond the scope*' as the minimal formulaic expression, while others may consider a larger span such as '*is beyond the scope of this paper*'. Here, we follow past research (Hirohata et al., 2008; Dayrell et al., 2012; Fiacco et al., 2019) and deal with this issue by regarding the whole sentence as the minimal unit of a communicative function.

Formulaic expressions and their communicative functions have been investigated mainly in academic writing research to help people write papers more rapidly and fluently (Cortes, 2013; Mizumoto et al., 2017; Omidian et al., 2018). There even exist some computer systems for academic-writing assistance[1][2] that rely on these communicative functions to improve the user's writing skills by suggesting commonly-used, alternative formulaic expressions. This is especially helpful for users whose native language is not English (Chen and Baker, 2010; AlHassan and Wood, 2015).

[1]http://langtest.jp/awsum/
[2]http://pep-rg.jp/abst/

Writing-assistance systems use pre-compiled lists of formulaic expressions labelled with communicative functions for each discipline. There are two approaches to create such lists (Biber et al., 2007): 1) the top-down approach, in which communicative functions of sentences are first identified and formulaic expressions are subsequently extracted from the sentences, and 2) the bottom-up approach, in which formulaic expressions are first extracted from a corpus and their communicative functions are subsequently identified. With either approach, problems arise when computational methods are applied to create the lists. For the top-down approach, no evaluation dataset is publicly available for classifying sentences into communicative functions. Moreover, evaluation datasets are expensive and time-consuming to build. To alleviate this issue, only smaller portions of papers, such as the abstract (Wu et al., 2006; Hirohata et al., 2008; Dayrell et al., 2012) or introduction (Pendar and Cotos, 2008), were annotated, and a limited number of disciplines were used (Cortes, 2013; Mizumoto et al., 2017). The bottom-up approach is not much better, because there is no established evaluation dataset for detecting formulaic expressions. Previous work, therefore, relied on domain experts to manually assess the quality of extracted formulaic expressions (Brooke et al., 2015; Iwatsuki and Aizawa, 2018), which, in addition to being costly, hinders replicability. Overall, the unavailability of annotated resources for both communicative functions and formulaic expressions has hindered the development of automated methods for detecting communicative functions.

There are, nonetheless, closely related resources for academic writing, in which examples of phrases and wordings are collected and classified into communicative functions. Academic Phrasebank (Morley, 2014) is one of them. However, the use of this resource as a ground-truth dataset is not

straightforward, as it was made with the purpose of helping scholars write and organise scientific papers. Therefore, it contains mostly incomplete sentences as example expressions (see Figure 1) and lacks the contextual information needed to detect communicative functions. Another problem with Academic Phrasebank is that example expressions were retrieved from papers belonging to a wide variety of disciplines ranging from humanities to medicine. Since section structures (Thelwall, 2019), vocabulary, word usage and the use of communicative functions differ among disciplines (Hyland, 2008), it is not reasonable to evaluate classifiers of communicative functions on that resource if one hopes to draw meaningful conclusions.

The present study attempts to address the aforementioned problems by building a new evaluation dataset for the detection of communicative functions of sentences. The proposed dataset contains unaltered, contextualised sentences collected from a domain-specific corpus, that is, the ACL Anthology Sentence Corpus (AASC)[3]. Sentences are annotated with communicative functions (and minimal formulaic expressions) by using a set of labels derived from Academic Phrasebank. We assume that the communicative function of a sentence can be realised at the sentence-level embedding. Accordingly, evaluating the accuracy of automatically detected communicative functions reduces to the evaluation of what is captured in sentence representations. Therefore, we propose a task of ranking sentence representations according to a given communicative function. We present a series of experiments comparing existing sentence representation models and show that contextualised, in-domain models (Beltagy et al., 2019) perform best at this task. Our results also show that state-of-the-art sentence representation models are still far behind human performance. This motivates the need for further investigation into not only semantic representations but also functional representations. The dataset is available on our GitHub repository (`https://github.com/Alab-NII/FECFevalDataset`).

Introduction Section

Stating the purpose of the current research

- The specific objective of this study was to …
- An objective of this study was to investigate …
- This thesis will examine the way in which the …
- This study set out to investigate the usefulness of …

…

Describing the research design and the methods used

- Data for this study were collected using …
- Five works will be examined, all of which …
- This investigation takes the form of a case-study of the …
- This study was exploratory and interpretative in nature.

…

Figure 1: Example expressions from Academic Phrasebank that are classified into communicative functions (written in bold).

[3] `https://github.com/KMCS-NII/AASC`

2. Related Work

2.1. Word or Phrase Suggestion for Academic Writing Assistance

When writing a research article, authors are often faced with a situation where they are not able to think of a desirable phrase to explain something or they wish to determine whether their wording is grammatically and conventionally correct. In such cases, they try to find better phrases or wordings by consulting books on academic writing or they search the web for phrases that appear more frequently. Because this process takes much time and effort, some computer systems have been proposed to automate this process. In most cases, phrases or wordings were extracted from linguistic resources and recorded in a database in advance, and a system searches for one of them based on the users' writing. In order to extract frequently used word n-grams, Jeong et al. (2014) relied on PubMed structured abstracts as a resource, in which sentences are labelled with the following functions: introduction, methods, results and discussion. However, this convention of writing abstracts is specific to PubMed; thus, this work will not be applicable to other disciplines. Chang and Chang (2015) and Yen et al. (2015) extracted grammatical phrase patterns from an English dictionary, rather than word n-grams from corpora. The system[4] they proposed is useful to find a correct usage of specific words. Liu et al. (2016) extracted frequent word n-grams from Elsevier's ScienceDirect and paraphrased them using WordNet synonyms to extend their database.

Despite the differences in the methods used to create databases, the method of recommendation of phrases and wordings is similar among the systems mentioned here. When a user writes something, all systems show examples or phrases that follow the user's input. For example, if a user writes '*We propose*', the systems only show phrases that contain '*propose*'. This is a limitation of keyword-based search in writing-assistance systems. In order to help users find phrases with different wordings, the use of communicative functions as queries, rather than keywords alone, can be beneficial.

2.2. Formulaic Expressions and Communicative Functions in Academic Prose

Because the usage of formulaic expressions has been found to differ across disciplines (Hyland, 2008), researchers have analysed the types of formulaic expressions used in scholarly papers in certain domains, including applied linguistics (Qin, 2014), social sciences (Lu et al., 2018), telecommunications (Pan et al., 2016), mathematics (Cunningham, 2017) and medicine (Jalali and Moini, 2014). Moreover, a few attempts have been made to create a list of formulaic expressions (Simpson-Vlach and Ellis, 2010; Ackermann and Chen, 2013).

However, it is necessary to categorise formulaic expressions in order for authors to use them efficiently. Simpson-Vlach and Ellis (2010) classified formulaic expressions into *functional categories*. Others focused on rhetorical

[4] `http://writeahead.nlpweb.org`

structure-based communicative functions as a category system. Swales (1981) introduced the concept of *move* as a rhetorical unit in the introduction section of research articles. Following his work, several investigations into the usage and transition of moves in scholarly papers were conducted, and it was found that patterns of moves are different across disciplines (Ozturk, 2007; Cotos et al., 2015; Maswana et al., 2015). Halliday and Matthiessen (2014) conducted broader analyses of functions in different levels of linguistic units ranging from multiple sentences to phrases. Furthermore, Lorés (2004) investigated the usage of Theme and Rheme in abstracts of scholarly articles.

2.3. Sentence Representations and Their Evaluation

Since the sentence is one of the fundamental units of languages, vector representations of sentences have attracted much research attention. Following successful word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), unsupervised methods to acquire sentence embeddings, such as Skip-Thought Vectors (Kiros et al., 2015) have been proposed. Conneau et al. (2017) found that even a supervised method trained on a dataset for natural language inference yielded universal sentence representations that perform well on various tasks. The current trend in the acquisition of sentence representations is the use of outputs from pre-trained language models such as BERT (Devlin et al., 2019).

In any case, sentence representations for general purposes do not always contain every aspect of languages. Hence, it is important to investigate which linguistic aspects they contain and comprehensive evaluation benchmarks have been proposed for this purpose (Conneau and Kiela, 2018; Wang et al., 2018). These benchmarks can well evaluate sentence representations in terms of semantic factors such as semantic relatedness, paraphrases and caption-image retrieval as well as logical factors such as entailment. Communicative functions, which the present paper focuses on, are another perspective related to rhetorical structure. Basically, the discourse structure is realised in multiple sentences, but a sentence can play the role of a rhetorical unit to make discourse. Therefore, rhetorical information embedded in sentence representation is worth evaluating.

3. Dataset Creation

3.1. Overview

This section describes the process we followed for building our dataset, which consists of sentences labelled with communicative functions. Figure 2 presents an illustration of this process. Starting from the example expressions provided in Academic Phrasebank, we queried a collection of scientific papers for candidate sentences, each of which was assigned to a communicative function. As most of the example expressions are domain dependent or too specific, we also performed an intermediate manual shortening step to generalise expressions and retrieve more sentences.

3.2. Academic Phrasebank

In the first step, we use Academic Phrasebank (Morley, 2014), which contains many example expressions labelled with communicative functions. An example is shown in Figure 1. Each example expression bears a formulaic expression, which is not explicitly marked. More than one thousand example expressions were collected and classified into 72 communicative functions (see Table 1). However, this resource has the two problems described in the introduction: incomplete sentences without context and expressions that are not domain specific. Therefore, it cannot be used as a ground-truth dataset.

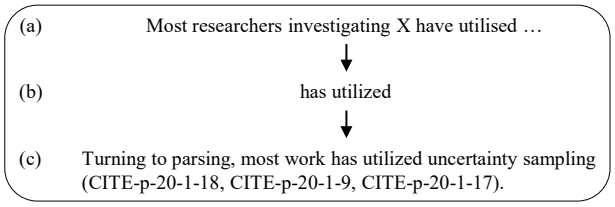

Figure 2: (a) Example expression collected in Academic Phrasebank, which is not a complete sentence. Most of the expressions do not appear in a corpus. Even formulaic expressions in the example expression are not used in a corpus because they are too long. (b) We choose a core formulaic expression (core FE) by shortening a formulaic expression. (c) By using the core FE as a query, we retrieve several sentences from a corpus.

Communicative functions are also modified because some are (1) based not on the rhetorical structure of a paper but rather on a grammatical perspective, (2) not distinguishable between each other or (3) not relevant for natural language processing (NLP), the discipline of the corpus we use. We present some examples here. Because of (1), '*Describing the process: infinitive of purpose*' and '*Describing the process: verbs used in the passive*' were integrated into one category named '*Describing the process*'. Because of (2), '*Reference to a previous investigation: researcher prominent*' and '*Reference to a previous investigation: investigation prominent*' were integrated. Because of (3), we removed the function '*Giving reasons for personal interest in the research*' as it is not common in the NLP community. After our modifications, the number of core formulaic expressions is 397, and the number of communicative functions is 39 (see Table 1).

3.3. Core Formulaic Expressions

We retrieve sentences from the corpus by using formulaic expressions as queries. Formulaic expressions are extracted from the example expressions by hand, but because they can be very specific or sometimes contain irrelevant content, some queries return no results. Therefore, we simplify and shorten the formulaic expressions and obtain what we call the core FEs to retrieve more sentences. For example, '*by adapting the procedure used by*' is a formulaic expression recorded in the resource, but it is not used in our corpus. Thus, we modify it to the core FE '*by adapting*'. The usage of core FEs causes noisy results; thus, we manually select sentences that have an intended communicative function after retrieving candidate sentences.

	Original		Modified	
	EEs	CFs	EEs	CFs
Introduction	328	17	104	11
Background	232	15	92	7
Method	210	14	82	6
Results	173	14	58	6
Discussion	153	12	61	9
Total	1,096	72	397	39

Table 1: Numbers of example expressions (EEs) and communicative functions (CFs) in Academic Phrasebank that we modified because many example expressions do not appear in the corpus and some communicative functions are not based on the rhetorical structure of scientific papers. We call the modified expressions the core FEs and only one core FE is annotated in each example expression.

3.4. Sentence Selection

We use the ACL Anthology Sentence Corpus (AASC) as our main source of sentences for several reasons. First, this dataset covers a limited range of disciplines that are all related to NLP, thereby standardising the usage of communicative functions and allowing us, as NLP researchers, to do annotation work. Second, each sentence in AASC is labelled with one out of five section headers (Introduction, Background, Method, Result and Discussion), which can be used to narrow down the number of possible communicative functions. To prevent research-topic-sensitive effects, all the sentences were retrieved from different papers in the corpus. Figure 3 shows a few instances in the proposed dataset. Each sentence has a sentence ID that corresponds to the sentence ID in AASC. Therefore, the surrounding context of each sentence can be easily retrieved if a classifier needs it.

Section: Introduction
Function: Limitation or lack of past work
Core FE: has not been investigated
Sentence: Also the extent to which inclusions pose a problem to existing NLP methods has not been investigated.
Sentence ID: D07-1016_s-2-1-0-3

Section: Result
Function: Reference to tables or figures
Core FE: figure * provides
Sentence: Figure 5 provides a more detailed characterization of LNQ's performance.
Sentence ID: P18-1029_s-12-6-1-0

Figure 3: Two examples recorded in the proposed dataset. Information on a section, communicative function and core FE is provided.

4. Quality Analysis of the Dataset

4.1. Method

In order to ensure that the sentence selection was correctly conducted and to assess the difficulty in detecting communicative functions, we performed manual evaluation for the dataset. Figure 4 shows the detailed design. Evaluators solved quizzes that were made from the dataset. In one quiz, three sentences are picked from a section in the dataset. One sentence is the targeted sentence and another sentence is the correct choice. Both have the same communicative function. The other sentence is the wrong choice (distractor) and has a different communicative function. The communicative function of the targeted sentences was given. Figure 5 shows an example of the quizzes.

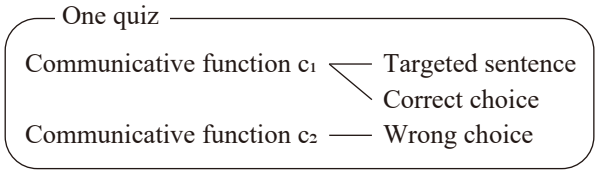

Figure 4: Design of the quizzes made from the dataset. The quiz consists of three sentences: a targeted sentence, correct choice and wrong choice. The targeted sentence and correct choice have the same communicative function (c_1), while the wrong choice has a different communicative function (c_2), which is not shown to evaluators.

Q: The purpose of this paper is to outline the main aspects of our ongoing and future work.
Function: The aim of the paper

(1) The aim of this paper is to deal with the first of these steps, i.e. question analysis module.

(2) This work uses a Maximum Entropy Markov Model (MEMM) based approach, which allows to combine different features.

Figure 5: Example of a quiz made from the dataset. The targeted sentence is denoted as Q. The communicative function of the targeted sentence is also shown. Evaluators are asked to choose a sentence that they think has the same communicative function out of (1) and (2). In this example, the answer is (1).

Each evaluator was asked to guess the communicative function of the sentences and choose the one that seemed to have the same communicative function as the targeted sentence. Because sentences are retrieved from different papers, the contents can be unrelated to each other, but the targeted sentence and the correct choice should be alike in terms of communicative functions. If an evaluator did not decide the answer, we did not include them as an evaluator for the quiz when calculating the accuracy. Four evaluators were assigned to Introduction and Background sections while five evaluators were assigned to the remaining sections (the different numbers of evaluators are coincidental).
After the evaluation, we calculated the accuracy and inter-evaluator agreement using Fleiss' κ. The accuracy indi-

cates how likely evaluators were to choose the correct answers, while the agreement indicates the degree to which they made the same choice. Thus, if the sentence selection in the process of creating the dataset fails to make pairs of sentences with the same communicative functions, the accuracy will be low but the agreement will be high. In other words, a low accuracy and high agreement indicate that the dataset is of low quality. In addition, if the task of detecting communicative functions is very difficult, evaluators will become confused, resulting in both a low accuracy and low agreement.

Section	# CF	# sent.	Acc.	κ
Introduction	11	104	97.9	93.0
Background	7	92	87.7	62.5
Method	6	82	78.4	40.7
Result	6	58	84.4	60.0
Discussion	9	61	85.2	60.7

Table 2: Numbers of sentences (# sent.) and communicative functions (# CF). # sent. and # CF are not balanced because the dataset is created based on Academic Phrasebank, which bears imbalance. The accuracy of annotators' choice (Acc.) and their agreement (κ, computed as Fleiss' Kappa) are also listed.

4.2. Results and Discussion

Table 2 presents the statistics of the dataset and the results. The accuracy and agreement in the table are macro averages of the accuracy and agreement for each communicative function. The results show that all the sections except *Method* yielded high accuracy and agreement, which implies that the dataset is of sufficient quality and the task is not too difficult. Communicative functions for *Introduction* yielded the highest scores even though the number of functions is higher than those of the others. This is probably because they do not overlap with each other. The *Method* section yielded a moderate accuracy and low agreement, which implies that the task is more difficult than the four other sections. Table 3 presents the confusion matrix for this section. The communicative function '*description of the process*' was found to be confused with others, probably because this communicative function is more general than the others. In other words, all sentences in *Method* could be labelled with that function. However, it is difficult to define communicative functions more finely for *Method* because methodology varies too widely among papers.

Table 5 lists the number of quizzes at different accuracy thresholds. We note that 64.7% of the data showed 100% accuracy, and the accuracy for 84.4% of the data is greater than 75%, which implies that the majority of the quizzes are easy to answer. Thus, the task of detecting the communicative functions of sentences is not too difficult for humans. It can also be said that communicative functions are understandable regardless of the content of a sentence. The accuracy is recorded in the dataset so that other researchers can use specific part of the data such as only data with 100% accuracy.

5. Evaluation of the Performance of Detection of Communicative Functions

5.1. Overview

In this section, we present two evaluation experiments to assess the performance of existing models for sentence representations in detecting communicative functions. First, we conducted the same task as mentioned in Section 4 to compare the performance of computational models to human performance. The second series of experiments adapts to a more realistic scenario and involves a ranking task, in which sentences are ordered according to a given communicative function based on their similarity.

5.2. Solving Quizzes with Computational Models

5.2.1. Task Description

Here, the task is to solve the quizzes described in the previous section by using existing sentence representation models. The cosine similarity between the targeted sentence and the two possible choices is used to determine the correct answer. Evaluation is performed by counting the number of correct choices that acquired higher similarity scores than the wrong choices.

We used two types of sentence representations: ones made from word embeddings and contextualised ones. As word-embedding-based sentence representations, we used Skip-gram (Mikolov et al., 2013)[5], GloVe (Pennington et al., 2014)[6] and Flair (Akbik et al., 2018)[7]. We first created word embeddings for each word in a given sentence with these models and then added them to one vector. We trained all word embeddings on AASC with punctuation marks removed.

We also used BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019) with their pre-trained models to create sentence representations. For BERT, we chose the `bert-large-uncased` pre-trained model. The pre-trained model for SciBERT is trained on scientific articles. We did not fine-tune the models, because the number of sentences in the dataset is too small.

5.2.2. Results and Discussion

Table 4 lists the results. Among the models we used, SciBERT yielded the best score. However, there is much room for improvement to approach human performance.

Furthermore, although humans were confused by the communicative functions in *Method*, some computational models did not always achieve the lowest score in that section. A comparison between BERT and SciBERT indicates that, even in the task of detecting communicative functions, training on the domain-specific corpus has a positive effect on the performance.

5.3. Baseline Performance of Ranking

5.3.1. Task Description

The ranking experiment consists of three steps. First, one sentence in the dataset is chosen as the targeted sentence. Second, all the other sentences in the same section are

[5] https://github.com/dav/word2vec
[6] https://nlp.stanford.edu/projects/glove/
[7] https://github.com/zalandoresearch/flair

	Communicative function	(1)	(2)	(3)	(4)	(5)	(6)
Answer	(1) Methodology used in past work	57		1	1	1	
	(2) Reasons why a method was adopted or rejected		55		1	5	4
	(3) Using methods used in past work	3		17			
	(4) Characteristics of samples or data			1	25	4	5
	(5) Criteria for selection	1	1		2	20	1
	(6) Description of the process	9	11	20	18	7	140

Table 3: Confusion matrix for the Method section. It is found that the communicative function '*description of the process*' is very confusing.

	Skip-gram			GloVe			Flair	BERT	SciBERT	Humans
dimension	200	200	500	200	200	500	2048	1024	768	N/A
window size	2	8	2	2	8	2	N/A	N/A	N/A	N/A
Introduction	68.3	64.4	65.4	57.7	60.6	57.7	67.3	59.6	84.6	97.9
Background	56.5	56.5	54.3	53.3	55.4	52.2	59.8	56.5	67.4	87.7
Method	52.4	51.2	54.9	53.7	54.9	52.4	57.3	48.8	59.8	78.4
Result	56.9	55.2	55.2	56.9	55.2	51.7	62.1	58.6	70.7	84.4
Discussion	55.7	54.1	55.7	60.7	54.1	59.0	50.8	55.7	62.3	85.2

Table 4: Comparison among the performance of the models we used and human performance. None of the models outperformed humans and SciBERT yielded the best performance among the sentence representations.

Accuracy (%)	100	≥ 75	≥ 50
Introduction	98(94%)	104(100%)	104(100%)
Background	61(66%)	78(85%)	90(98%)
Method	30(37%)	57(70%)	77(94%)
Result	33(57%)	45(78%)	53(91%)
Discussion	35(57%)	51(84%)	57(93%)
All	257(65%)	335(84%)	381(96%)

Table 5: Distribution of the quizzes in terms of the accuracy. 64.7% of the dataset showed 100% accuracy.

	SG	GloVe	Flair	BERT	SciB
Introduction	32.7	30.6	33.2	32.6	48.2
Background	35.9	33.8	39.4	37.3	44.9
Method	42.2	40.0	40.3	42.6	49.5
Result	47.7	44.3	48.2	43.6	57.7
Discussion	26.3	24.9	24.1	24.2	32.9

Table 6: Mean average precisions for Skip-gram (SG), GloVe, Flair, BERT and SciBERT (SciB). For Skip-gram, the dimension is 200 and the window size is 2, and for GloVe, the dimension is 200 and the window size is 8, which are the best values in the experiment.

Target: Although CG is a radically lexicalist grammatical theory, little attention has been paid to the structure of the lexicon.

#	Sentences	Cosine	Correct
[1]	Recently there has been interest in the development of a general computational treatment of the comparative.	0.9046	
[2]	Dependency parsing is a basic technology for processing Japanese and has been the subject of much research.	0.8974	
[3]	Although it has been suggested that head-driven parsing has benefits for lexicalist grammars, this has not been established in practice.	0.8955	✓
[4]	While it has been observed informally that the internal sentence representations of such models can reflect semantic intuitions (CITE-p-15-4-3), it is not known which architectures or objectives yield the 'best' or most useful representations.	0.8820	✓
[5]	Below, it will be argued that these semantic representations are indeed too weak, but not only from the point of view of Natural Language Processing.	0.8801	

Figure 6: Illustration of the ranking task. Cosine similarities between a targeted sentence and all the other sentences in its section are calculated, and sentences are ranked by the similarity score. The sentences that have the same communicative function as the targeted sentence are marked correct. In this example, sentences 3 and 4 have the same communicative function.

ranked in the order of their cosine similarity with the targeted sentence (Figure 6). Third, we calculate the mean average precision, which is described below, by checking whether each ranked sentence has the same communicative function as the targeted sentence. These steps were applied to all the sentences in the dataset, after which the average of all the scores was calculated for each section. Sentence representations were made in the same manner as in the previous subsection.

We calculate the average precision for all targeted sentences and average them to obtain the mean average precision (MAP). The average precision is a metric often used for information retrieval (Manning and Schütze, 1999) and is calculated using the following steps. First, the precision at each rank is calculated from the top. Then, the average of the ranks, called average precision, is calculated. MAP is formulated as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{n_i}\frac{c_j^i}{j},$$

where n is the number of targeted sentences, m is the number of sentences to be compared with a targeted sentence and c_j^i is the number of correct choices contained in the ranked list from the top to the j-th one.

5.3.2. Results and Discussion

Table 6 lists the results. We observe that the models perform rather poorly on the ranking task. The models show the same tendency as in the previous experiments; that is, the highest performance is achieved by SciBERT and the results for the *Method* section are not significantly lower than for the other sections. On the other hand, the *Results* section yielded the highest score in this experiment. The difficulty in the *Introduction* section for computational models results from the greater number of communicative functions; the Introduction section has approximately twice as many communicative functions as the *Method* section. Moreover, the vocabulary size is different between the two sections: 1,238 for *Introduction* and 853 for *Method*. The *Introduction* section contains varying contexts. In other words, humans are not confused with many classes and a large vocabulary in one section because they can focus on formulaic expressions in sentences and recognise communicative functions correctly. This implies that current computational models mostly rely on content in a sentence and fail to elicit communicative functions.

These results can be used as baselines for future detection methods of communicative functions.

6. Conclusion

We proposed a new evaluation dataset to check whether sentence representations can detect communicative functions of sentences. We performed comparisons of recent models on the dataset and found that SciBERT, which was trained on a corpus of research articles, performed best.

Further investigation should be conducted into sentence representations not only from a semantic perspective but also from a functional perspective. Each sentence in our dataset has a reference to AASC; therefore, even a whole document can be used to apply machine-learning models to classification, if necessary.

Our approach to create evaluation datasets can be applied to corpora of other domains as well. The core FEs may also serve as helpful queries for the retrieval of sentences from other corpora.

7. Acknowledgements

This research is supported by JSPS KAKENHI Grant Numbers 19J12466 and 18H03297 and by Atlanstic 2020 sabbatical grant IKEBANA.

8. Bibliographical References

Ackermann, K. and Chen, Y.-H. (2013). Developing the academic collocation list (ACL) — a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4):235 – 247.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

AlHassan, L. and Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting l2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, 17:51–62.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing*, pages 3606–3611.

Biber, D., Connor, U., and Upton, T. A. (2007). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structures*. John Benjamins Publishing.

Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., and Shein, F. (2015). Building a lexicon of formulaic language for language learners. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 96–104.

Chang, J. and Chang, J. (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.

Chen, Y.-H. and Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2):30–49.

Conneau, A. and Kiela, D. (2018). SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

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.

Cotos, E., Huffman, S., and Link, S. (2015). Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes*, 19:52 – 72.

Cunningham, K. J. (2017). A phraseological exploration of recent mathematics research articles through key phrase frames. *Journal of English for Academic Purposes*, 25:71 – 83.

Dayrell, C., Candido, A. J., Lima, G., Machado, D. J., Copestake, A., Feltrim, V., Tagnin, S., and Aluisio, S. (2012). Rhetorical move detection in English abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the Eight Interna-*

tional Conference on Language Resources and Evaluation (LREC'12), pages 1604–1609.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.

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

Fiacco, J., Cotos, E., and Rosé, C. (2019). Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 310–319.

Halliday, M. A. K. and Matthiessen, C. M. (2014). *Halliday's Introduction to Functional Grammar*. Routledge.

Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 381–388.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Iwatsuki, K. and Aizawa, A. (2018). Using formulaic expressions in writing assistance systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2678–2689.

Jalali, Z. S. and Moini, M. R. (2014). Structure of lexical bundles in introduction section of medical research articles. *Procedia - Social and Behavioral Sciences*, 98:719 – 726. Proceedings of the International Conference on Current Trends in ELT.

Jeong, S., Nam, S., and Park, H.-Y. (2014). An ontology-based biomedical research paper authoring support tool. *Science Editing*, 1(1):37–42.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Liu, Y., Wang, X., Liu, M., and Wang, X. (2016). Writerighter: An academic writing assistant system. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4373–4374.

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

Lu, X., Yoon, J., and Kisselev, O. (2018). A phrase-frame list for social science research article introductions. *Journal of English for Academic Purposes*, 36:76 – 85.

Manning, C. D. and Schütze, H., (1999). *Foundations of Statistical Natural Language Processing*, page 535. MIT Press.

Maswana, S., Kanamaru, T., and Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand*, 2:1–11.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

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.

Morley, J. (2014). Academic Phrasebank. `http://www.phrasebank.manchester.ac.uk/`.

Omidian, T., Shahriari, H., and Siyanova-Chanturia, A. (2018). A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for Academic Purposes*, 36:1 – 14.

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

Pan, F., Reppen, R., and Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes*, 21:60 – 71.

Pendar, N. and Cotos, E. (2008). Automatic identification of discourse moves in scientific article introductions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–70.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System*, 42:220 – 231.

Simpson-Vlach, R. and Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4):487–512.

Swales, J. M. (1981). *Aspects of Article Introductions*. The University of Michigan Press.

Swales, J. M. (1990). *Genre Analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. (2004). *Research Genres: Explorations and Applications*. Cambridge University Press.

Thelwall, M. (2019). The rhetorical structure of science? a multidisciplinary analysis of article headings. *Journal of Informetrics*, 13(2):555 – 563.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Wu, J.-C., Chang, Y.-C., Liou, H.-C., and Chang, J. S.

(2006). Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 41–44.

Yen, T.-H., Wu, J.-C., Chang, J., Boisson, J., and Chang, J. (2015). WriteAhead: Mining grammar patterns in corpora for assisted writing. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 139–144.