Reusable Phrase Extraction Based on Syntactic Parsing

Xuemin Duan[1]**, Hongying Zan***[1]**, Xiaojing Bai**[2] **and Christoph Zahner**[3]

[1] School of Information Engineering, Zhengzhou University, Zhengzhou, China
[2] Language Centre, Tsinghua University, Beijing, China
[3] University of Cambridge Language Centre, UK

xueminduan@163.com, iehyzan@zzu.edu.cn, bxj@tsinghua.edu.cn, cz201@cam.ac.uk

Abstract

Academic Phrasebank is an important resource for academic writers. Student writers use the phrases of Academic Phrasebank organizing their research article to improve their writing ability. Due to the limited size of Academic Phrasebank, it can not meet all the academic writing needs. There are still a large number of academic phraseology in the authentic research article. In this paper, we proposed an academic phraseology extraction model based on constituency parsing and dependency parsing, which can automatically extract the academic phraseology similar to phrases of Academic Phrasebank from an unlabelled research article. We divided the proposed model into three main components including an academic phraseology corpus module, a sentence simplification module, and a syntactic parsing module. We created a corpus of academic phraseology of 2,129 words to help judge whether a word is neutral and general, and created two datasets under two scenarios to verify the feasibility of the proposed model.

Keywords: Academic Phraseology Extraction·Academic Phrasebank·Syntactic Parsing.

1 Introduction

The Academic Phrasebank is a general resource created by the University of Manchester for academic writers. And the items of it are neutral and generic, which means that you don't have to worry about accidentally stealing someone else's idea when using these items in your academic paper. The Reusable phrases including the phrases in Academic Phrasebank do not have a unique or original construction, not express a special point of view of another writer.

Now, most of the assisted academic writing research focused on automated essay scoring (AES), but different from the ordinary essay prefer to life and social, research article is a scientific record of scientific research result or innovation thinking in theoretical, predictive, and experimental. Research article writing has more rigorous grammar, discourse structure and phraseology. (Davis and Morley, 2018) mentioned that the central of designing teaching activities developed by Academic Phrasebank is the purpose of improving the cognitive ability of student writers to potential plagiarism. Learning academic phraseology can effectively help student writers avoid plagiarism, and student writers can use the learned academic phraseology in their own research article writing, so as to improve their academic writing ability. However, Academic Phrasebank does not cover all academic phraseology in authentic research articles, so it is necessary to extract academic phraseology automatically from more unlabelled text to expand our "Academic Phrasebank".

Plenty of research relating to teaching activities about Academic Phrasebank, but little or nothing that concerns extracting academic phraseology automatically. Therefore, in this paper, we introduce an Academic phraseology extraction model based on constituency parsing and dependency parsing, which aims to extract similar samples with phrases of Academic Phrasebank from unlabelled research articles. The academic phraseology examples are shown in Table 1.

In order to analyze the semantics and structure of unlabelled sentences, we first create a corpus of academic phraseology which including all words of the phrases of Academic Phrasebank. Due to the items of Academic Phrasebank all are general and neutral, the word in a given sentence which also belongs to the corpus of academic phraseology can exist in the extraction result.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1166-1171,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

As there is no relevant study at present, this paper did not select others' baseline for comparison but created two datasets under two scenarios to verify the feasibility of the proposed model. A dataset is completed from the phrases of Academic Phrasebank, which is more standard, while a dataset is annotated from authentic research articles with more complex sentence structure, and the experimental results demonstrate the different effectiveness of the proposed model on a different dataset.

In brief, the main contributions are as follows:

– We propose a new task, named Academic Phraseology Extraction, which contributes to academic writing and provides valuable phrases for student writers to organise their research articles.

– We propose a model by syntactic parsing for Academic Phraseology Extraction, which considers phrase structure, dependency and semantic analysis of the given sentence.

– We collect sentences from authentic research articles and construct a dataset for Academic Phraseology Extraction with human-annotation. In addition, we also collect phrases from Academic Phrasebank and construct a dataset for Academic Phraseology Extraction with human-completion.

Sentences	**Academic Phraseology**
This paper have argue that the proposed TDNN could be further improved.	This paper have argue that ...
There have been efforts in developing AES approaches based on DNN.	There have been efforts in developing ...
Further study are required to identify the effectiveness of proposed AES.	Further study are required to identify the effectiveness of ...

Table 1: Academic Phraseology Extraction Results

2 Related Work

Corpus of contemporary American English (COCA) is the latest contemporary corpus of 360 million words developed by (Davis, 2008). It covers five types of the corpus of novels, oral English, popular magazines, and academic journals in different periods in the United States. Using COCA to study can make up for the lack of students' understanding of vocabulary, and at the same time, it can cultivate favorable conditions for essay writing. However, the COCA is inappropriate to be used as a corpus for judging whether a word belongs to academic phraseology in the process of academic phraseology extraction. So we create a corpus of academic phraseology for this paper.

Academic Phrasebank is a general resource developed by Dr. John Morley of the University of Manchester to help student writers writing. (Davis and Morley, 2018) has designed some relevant teaching activities developed based on Academic Phrasebank. The research holds that the most important two points of academic writing teaching purpose are to obtain timely writing feedback and improve the cognitive ability of student writer to plagiarism in academic writing. The former means automated research article scoring, and the latter means strengthening students' learning of phrases in Academic Phrasebank and authentic research article. Because the content of Academic Phrasebank is neutral and general, frequent learning of Academic Phrasebank can help students improve their cognitive ability. But the content of Academic Phrasebank is limited. If student writer want to expand their own "Academic Phrasebank", they need to extract academic phraseology from authentic research articles.

The problem of analyzing complex sentences in natural language processing is to make sentences simple to understand, by identifying clause boundaries. Before extracting academic phraseology from a sentence, we choose to simplify the sentence first. (Sharma, 2016) provides a survey of predicting clause boundaries while. (Sacaleanu, 2017) proposed a rule-based method for clause boundary detection. The latter method is a pipeline that uses phrase structure trees to determine the clauses.

3 Our Approach

In this section, we will introduce our academic phraseology extraction approach. There are three main components in our model, i.e., an academic phraseology corpus module to help identify whether a word in a sentence belongs to academic phraseology, a sentence simplification module to prevent incomplete academic phraseology from being extracted, and a syntactic parsing module to determine the final results of academic phraseology extraction. We will introduce the details of our academic phraseology extraction approach as follows.

3.1 Corpus of Academic Phraseology

The academic phraseology extraction is extracting based on the dependency and constituency structure of a sentence, but the final extraction results of two sentences composed of the same dependency and constituency structure are not necessarily the same, because the content of academic phraseology is also related to the semantics of words of a sentence. For example, there are two sentences that only have different subjects, ”Further study” and ”Bert and transformer”. Although they both act as the components of nominal phrases in the sentences, the former can appear in the result, but the latter can not. This is because the content of ”Further” and ”study” are all neutral and general, but ”Bert”, ”and” and ”transformer” have a special word. How to judge whether a word is neutral and general? we need a corpus containing a large number of neutral and general words to help us judge.

The corpus of academic phraseology we created contains all words in Academic Phrasebank, which helps us judge whether a word or phrase should appear in the final result. It has a pivotal role in extracting academic phraseology from the unlabelled text. As the phrases in Academic Phrasebank are all academic phraseology, In the process of academic phraseology extraction, the words of a sentence that appear in Academic Phrasebank can all appear in the result of academic phraseology extraction.

We segmented the phrases in Academic Phrasebank and deleted the repeated words to obtain the corpus of academic phraseology. Academic Phrasebank contains 12,451 phrases, and the resulting corpus of academic phraseology we constructed contains 2,129 words.

3.2 Sentence Simplification

English sentences are mainly composed of subject, predicate, object, attribute, adverbial, complement, and other components, in which the predicate component can only be composed of verbs, and the rest of the sentence components can be composed of words or replaced by clauses. English sentences containing clauses are often long and complex, and it is difficult to extract academic phraseology from them. Therefore, for complex sentences, it is necessary to divide them into simple clauses first.

The sentence simplification is to identify more than two English sentences with more than two clauses, mark the boundary of the clauses, and decompose the complex sentences into many simple sentences. In order to improve the accuracy of academic phraseology extraction, we first simplify the complex sentences before extraction and then extracts the academic phraseology from the simple sentences. This kind of syntactic text simplification is non-destructive. It mainly extracts embedded clauses from sentences with complex structures, so as to rewrite them without affecting their original meanings. This process reduces the average sentence length and complexity, making the text simpler. The key point of sentence simplification is to extract the implied clause from the sentence with a complex structure

In this paper, we identify the relationship between the main sentence and the paratactic or subordinate sentence by constituency parsing, classify the subordinate sentence, determine the optimal clause boundary in the sentence, and extract the clause from the constituency parse tree by using the defined rules.

First, get a constituency parse tree of given complex English sentence, then identify the non-root clausal node of the constituency parse tree (e.g. SBAR, S.) and remove it from the main tree but retain these subtrees, then remove all hanging in the main tree prepositions, subordinate conjunctions and adverbs, the result was simplified sentences. The sentence simplification examples are shown in Table 2.

Sentences	**Simplified Sentences**
The prompt-dependent models can hardly learn generalized rules from rated essays for nontarget prompts, and are not suitable for the prompt independent AES.	[”The prompt-dependent models can hardly learn generalized rules from rated essays for nontarget prompts.”, ”The prompt-dependent models are not suitable for the prompt independent AES.”]
A supervised model is employed to identify the essays in a given set of essays, and it aims to recognize the essays with the extreme quality in the test dataset.	[”A supervised model is employed to identify the essays in a given set of essays.”, ”A supervised model aims to recognize the essays with the extreme quality in the test dataset.”]
Such relative precision is at least 80% on different prompts so that the overlap of the selected positive and negative essays is fairly small.	[”Such relative precision is at least 80% on different prompts.”, ”The overlap of the selected positive and negative essays are fairly small.”]

Table 2: Sentence Simplification Results

3.3 Syntactic Parsing

Our academic phraseology extraction approach is a rule-based approach using constituency parse tree and dependency tree. By identifying the main verb and determining which nominal phrases of the sentence belongs to academic phraseology by the corpus of academic phraseology, we can easily extract the academic phraseology from the sumolified sentence.

The steps for extracting academic phraseology are explained with the help of the following examples: ”Further study are required to identify the effectiveness of poposed AES.”

Step 1: Obtaining the dependency tree of the given simplified sentence to identify the main verb. The dependency tree is shown in Figure 1, we can get the main verb is ”required”.

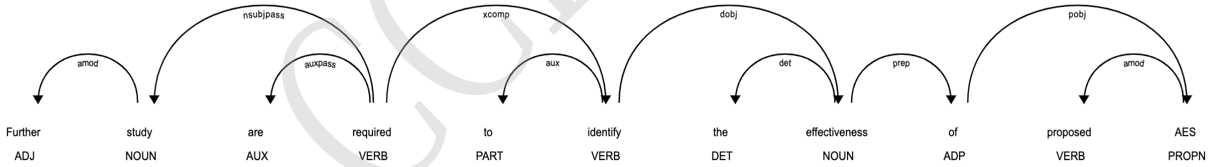

Figure 1: Dependency Parse Tree

Step 2 Obtaining the constituency parse tree to identify all nominal phrases in a sentence and their order. The constituency parse tree is shown in Figure 2.

Step 3 Taking the main verb as the center and classifying the nominal phrases with left part of verb or right part of verb. Then, using the corpus of academic phraseology to determine whether a nominal phrase is deleted or retained. If the left part of the main verb occupied in the corpus of academic phraseology means that it can be retained. The right part of the main verb is divided into several noun phrases and analyzed from the first one. If the first one belongs to the corpus of academic phraseology, then continue to analyze the next one. If not, delete it and the part on its right,and then finish the analysis. All nominal phrases of this sentence and their determines are shown in Table 3.

The first nominal phrase, ”Further study”, can be retained because ”further” and ”study” all exist in the corpus of academic phraseology. So is the second nominal phrase. The third nominal phrase, ”proposed AES”, should be deleted since ”AES” not exist in the corpus of academic phraseology. According to Table 3, we can get the result of academic phraseology extraction is ”Further study are required to identify the effectiveness of ...”

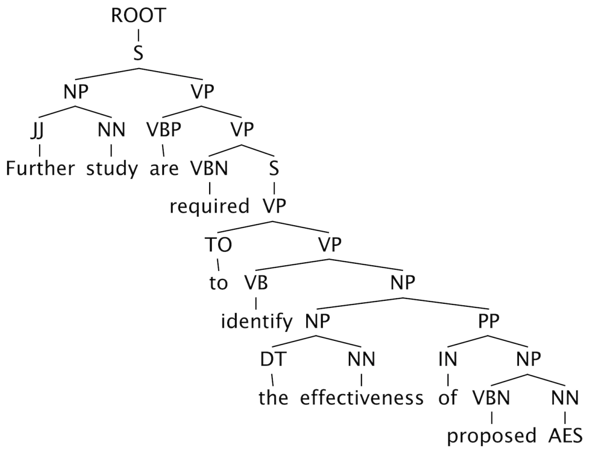

Figure 2: Constituency Parse Tree

	Nominal Phrases	**Judgements**
left part	Further study	retain
right part	the effectiveness	retain
	proposed AES	delete

Table 3: Nominal Phrases Judgements.

4 Experiments

In this section, we present our experiment datasets and results, which devote to answering the following questions that how effective is the proposed academic phraseology extraction model in extracting academic phraseology from sentences written according to the phrases in Academic Phrasebank and whether this model can obtain a same performance in extracting academic phraseology from real academic papers compared to the former.

4.1 Datasets

Since there are not existing academic phraseology dataset now,we created two datasets under two scenarios, ”standard” and ”authentic”, to verify the feasibility of the proposed model. The ”standard” dataset is completed from the phrases of Academic Phrasebank by human. There is no special sentence pattern in sentences completed from Academic Phrasebank,which means that this dataset is more standard. The ”authentic” dataset is annotated from authentic research articles by human. There are some special sentence patterns in sentences of authentic research article, which means that this dataset contains many complex sentence patterns that may appear in authentic research paper, such as inverted sentences and accent sentences. It is more ”authentic”.

We took 1,000 phrases from academic phrasebank and manually completes them into sentences. In addition, we also selected 1,000 complete sentences from authentic research articles and manually annotate their academic phraseology. They are combined together to form the academic phraseology datasets in this paper. The contents are shown in the Table 4.

4.2 Evaluation Metrics

In the process of extracting academic phraseology from sentence, we hope to get more words of our predicted academic phraseology that are the same as those in true academic phraseology. Based on this sense, we calculate Precision, Recall and F score for academic phraseology extraction model.

Datasets	Sentences
Academic Phrasebank Phraseology Dataset	1,000
Authentic Research Articles Phraseology Dataset	1,000

Table 4: Academic Phraseology Extraction Model Datasets.

4.3 Results and Analysis

We use the proposed academic phraseology model to experiment with two datasets, the overall experimental results are shown in Table 5.

From the overall results, we can observe that the performance of the proposed model on Academic Phrasebank Phraseology Dataset is better than on Authentic Research Articles Phraseology Dataset. This is because the academic phraseology extraction model proposed in this paper is designed for the common sentence pattern with the highest frequency in research articles. The Authentic Research Articles Phraseology Dataset has more special sentence patterns, such as inverted sentences and accent sentences.

There is still a lot of room for improvement. If we analyze and modify the proposed academic phraseology extraction model separately for the special sentence patterns that appear less frequently in research articles, the performance of the proposed model on all datasets will be improved.

Datasets	Precision	Recall	F1 score
Academic Phrasebank Reusable Phrases Dataset	0.96	0.62	0.72
Authentic Research Articles Reusable Phrases Dataset	0.84	0.99	0.88

Table 5: The Performance of Reusable Phrase Extraction Model on Different Datasets.

5 Conclusion

In this paper, we define a new task in assisted writing, Academic Phraseology Extraction, which devotes to providing valuable phrases for student writers to write their research articles. For extracting the similar samples with the phrases of Academic Phrasebank, we proposed an academic phraseology extraction model. The proposed model are divided into three components: corpus of academic phraseology, sentence simplification and syntactic parsing. Experiments on a academic phrasebank phraseology dataset and a authentic research article phraseology dataset validate the effectiveness of our approach.

References

Davis, M. and Morley, J. 2018. *Journal of Learning Development in Higher Education ISSN*, p.667X.

Davies, M. 2008. *The corpus of contemporary American English: 450 million words*, 1990-present.

Sharma, S.K. 2016. *International Journal of Computer Applications & Information Technology*,8(2), p.152.

Sacaleanu, B., Marascu, A. and Jochim, C. 2017. *International Business Machines Corp*,U.S. Patent 9,652,450.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D. 2014. *The Stanford CoreNLP natural language processing toolkit*,(pp.55-60).

Oakey, D. 2020. *Journal of English for Academic Purposes*,44, p.100829.