Reusable Phrase Extraction Based on Syntactic Parsing

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

[1] School of Information Engineering, Zhengzhou University, Zhengzhou, China
xueminduan@163.com, iehyzan@zzu.edu.cn
[2] Language Centre, Tsinghua University, Beijing, China
bxj@tsinghua.edu.cn
[3] University of Cambridge Language Centre, UK
cz201@cam.ac.uk

Abstract. Academic Phrasebank is an important resource composed of neutral and generic phrases for academic writers. In this paper, we name these neutral and generic phrases reusable phrases, and student writers use them to organize their research articles. Due to the limited size of Academic Phrasebank, it can not meet all the academic writing needs. There are still a large number of reusable phrases in authentic research articles. In order to make up for the deficiency of Academic Phrasebank, we proposed a reusable phrase extraction model based on constituency parsing and dependency parsing to automatically extract reusable phrases from unlabelled research articles. We divided the proposed model into three main components including a reusable words corpus module, a sentence simplification module, and a syntactic parsing module. We created a reusable words corpus of 2129 words to help judge whether a word is neutral and generic, and created two datasets under two scenarios to verify the feasibility of the proposed model.

Keywords: Reusable Phrase 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 all 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. In this paper, we name these neutral and generic phrases reusable phrases. Reusable phrase means that student writers can use these phrases in their paper without worrying about plagiarism due to they are neutral and generic. 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. [1] 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 reusable phrases can effectively help student writers avoid plagiarism, and student writers can use the learned reusable phrases in their own research article writing, so as to improve their academic writing ability. However, Academic Phrasebank does not cover all reusable phrases in authentic research articles. In order to make up for the deficiency of Academic Phrasebank, we propose a new task named reusable phrase extraction. The purpose of this task is to automatically extract reusable phrases from unlabelled research articles, so as to help writers build their own "Academic Phrasebank".

Plenty of research relating to teaching activities about Academic Phrasebank, but little or nothing that concerns extracting reusable phrases automatically. Therefore, in this paper, we introduce a reusable phrase 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 reusable phrases examples are shown in Table 1.

Table 1. Academic Phraseology Extraction Results.

Sentences	Reusable Phrases
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 ...

In order to analyze the semantics and structure of unlabelled sentences, we first create a reusable words corpus 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 reusable words corpus can exist in the extraction result.

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 Reusable Phrase 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 Reusable Phrase 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 Reusable Phrase Extraction with human-annotation. In addition, we also collect phrases from Academic Phrasebank and construct a dataset for Reusable phrases Extraction with human-completion.

2 Related Work

Corpus of contemporary American English (COCA) is the latest contemporary corpus of 360 million words developed by [2]. 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 reusable phrases in the process of reusable phrases extraction. So we create a reusable phrases corpus for this paper.

Academic Phrasebank is a general resource developed by Dr. John Morley of the University of Manchester to help student writers writing.[1] 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 reusable phrases 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 reusable phrases from a sentence, we choose to simplify the sentence first. [3] provides a survey of predicting clause boundaries while. [4] 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 reusable phrases extraction approach. There are three main components in our model, i.e., a reusable words corpus module to help identify whether a word in a sentence belongs to reusable phrases, a

sentence simplification module to prevent incomplete reusable phrase from being extracted, and a syntactic parsing module to determine the final results of reusable phrases extraction. We will introduce the details of our reusable phrases extraction approach as follows.

3.1 Reusable Words Corpus

The reusable phrases extraction model 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 reusable phrases 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, reusable phrases, to help us judge.

The reusable words corpus 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 reusable phrases from the unlabelled text. As the phrases in Academic Phrasebank are all reusable phrases, In the process of reusable phrase extraction, the words of a sentence that appear in Academic Phrasebank can all appear in the result of reusable phrases extraction.

We segmented the phrases in Academic Phrasebank and deleted the repeated words to obtain the reusable words corpus. Academic Phrasebank contains 12,451 phrases, and the resulting reusable words corpus 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 reusable phrases 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 reusable phrase extraction, we first simplify the complex sentences before extraction and then extracts the reusable phrases 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.

Table 2. Sentence Simplification Results.

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.”]

3.3 Syntactic Parsing

Our reusable phrase 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 reusable phrases by the reusable words corpus, we can easily extract the reusable phrases from the simplified sentence.

The steps for extracting reusable phrases are explained with the help of the following examples : “Further study are required to identify the effectiveness of proposed 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".

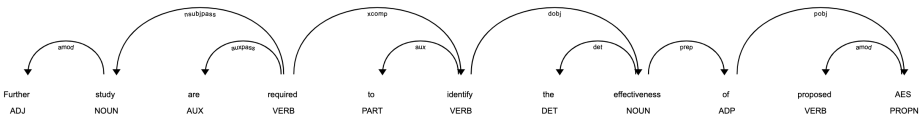

Fig. 1. Dependency Parse Tree.

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.

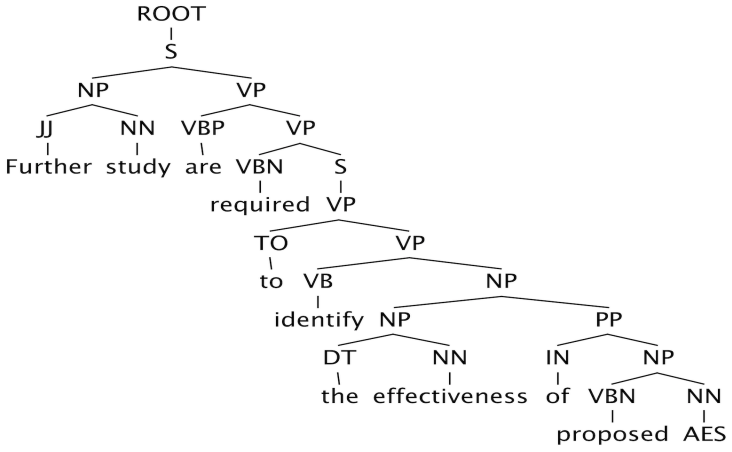

Fig. 2. Constituency Parse Tree.

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 reusable words corpus to determine whether a nominal phrase is deleted or retained. If the left part of the main verb occupied in the reusable words corpus 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 reusable words corpus, 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 reusable words corpus. So is the second nominal

Table 3. Nominal Phrases Judgements.

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

phrase. The third nominal phrase, "proposed AES", should be deleted since "AES" not exist in the reusable words corpus. According to Table 3, we can get the result of reusable phrase extraction is "Further study are required to identify the effectiveness of ..."

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 reusable phrase extraction model in extracting reusable phrase from sentences written according to the phrases in Academic Phrasebank and whether this model can obtain a same performance in extracting reusable phrase from real academic papers compared to the former.

4.1 Datasets

Since there are not existing reusable phrase 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 reusable phrases. They are combined together to form the reusable phrase datasets in this paper. The contents are shown in the Table 4.

4.2 Evaluation Metrics

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

Table 4. Reusable Phrase Extraction Model Datasets.

Datasets	Sentences
Academic Phrasebank Reusable Phrases Dataset	1,000
Authentic Research Articles Reusable Phrases Dataset	1,000

4.3 Results and Analysis

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

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

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

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 reusable phrase 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.

5 Conclusion

In this paper, we define a new task in assisted academic writing, Reusable Phrase Extraction, which devotes to providing valuable phrases for student writers to write their research articles. Learning the reusable phrases in the Academic Phrasebank is an important part of improving writing ability. For extracting the similar samples with the phrases of Academic Phrasebank, we proposed a reusable phrase extraction model. The proposed model are divided into three components: reusable words corpus, sentence simplification and syntactic parsing. Experiments on an Academic Phrasebank Reusable Phrases Dataset and an Authentic Research Article Reusable Phrases Dataset validate the effectiveness of our approach.

References

1. Davis, M., Morley, J.: Facilitating learning about academic phraseology: teaching activities for student writers. Journal of Learning Development in Higher Education (2018)
2. Davis, M.: The corpus of contemporary American English: 450 million words, 1990-present. (2008)
3. Sharma, SK.: Clause Boundary Identification for Different Languages: A Survey. International Journal of Computer Applications & Information Technology **8**(2), p.152(2016)
4. Sacaleanu, B., Marascu, A., Jochim, C.: Rule-based syntactic approach to claim boundary detection in complex sentences. International Business Machines Corp U.S. Patent 9,652,450 (2017)
5. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations 55–60(2014)
6. Oakey, D.: Phrases in EAP academic writing pedagogy: Illuminating Halliday's influence on research and practice. Journal of English for Academic Purposes **44**, 100829(2020)