Automatic Extraction of Phrasal Expressions for Supporting English Academic Writing

Shunsuke Kozawa, Yuta Sakai, Kenji Sugiki and Shigeki Matsubara

Abstract English academic writing is not easy for non-native researchers. They often refer to lexica of phrases on English research papers to know useful expressions in academic writing. However, lexica on sales do not have enough amount of expressions. Therefore, we propose a method for automatically extracting useful expressions from English research papers. We found four characteristics of the expressions by analyzing the existing lexicon of phrases on English research papers. The expressions are extracted from research papers based on statistical and syntactic information. In our experiment using 1,232 research papers, our proposed method achieved 57.5% in precision and 51.9% in recall. The f-measure was higher than the baselines, and therefore, we confirmed the feasibility of our method.

1 Introduction

The aim of our research is to support English academic writing because it is not an easy task for non-native researchers although English academic writing is indispensable for researchers to present their own research achievement. The researchers

Shunsuke Kozawa
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan, e-mail: kozawa@el.itc.nagoya-u.ac.jp

Yuta Sakai
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan, e-mail: sakai@el.itc.nagoya-u.ac.jp

Kenji Sugiki
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan, e-mail: sugiki@el.itc.nagoya-u.ac.jp

Shigeki Matsubara
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan, e-mail: matubara@nagoya-u.jp

often consult bilingual dictionaries to translate source language words into English words, refer to lexica of phrases on English research papers to know useful expressions in academic writing, and use search engines to learn English grammar and usage.

Some studies for supporting English writing have been conducted by focusing on English grammar and usage. Search systems for example sentences [5, 7, 13, 4, 2] and automatic correction systems [6] have been developed to assist confirmation of English grammar and usage. In contrast, no studies focusing on useful expressions in academic writing have been conducted. Researchers use lexica on sales (e.g. [10, 12]) to find the expressions. The lexica are useful because researchers can use expressions in them without any modification. However, the lexica do not have enough amount of expressions and example sentences because they are produced manually. Therefore, researchers have to refer to research papers in their fields to search for the expressions if they could not find appropriate expressions in the lexica.

In this issue, if lexica of useful expressions in academic writing could be automatically generated, they would help researchers to write research papers. Recently, a large amount of research papers have been published electrically [11]. This allows us to create a lexicon specialized in a particular field.

This paper proposes a method for automatically extracting useful expressions from English research papers. We call the useful expression *phrasal expression*. The phrasal expressions, which include idioms, idiomatic phrase, and collocations, are extracted from research papers using statistical and syntactic information.

2 Characteristics of Phrasal Expression

Phrasal expressions are useful expressions in academic writing and include idioms, idiomatic phrases and collocations. Table 1 shows examples of phrasal expressions. In order to capture characteristics of phrasal expressions, we analyzed the expressions appeared in the book [10], which is one of the most frequently used books to write English research papers. By analyzing of 1,119 expressions appeared in the book, we found four characteristics of phrasal expressions; a unit of phrasal expressions, phrasal sign (see Sec. 2.2), statistical characteristics and syntactic constraints. The following subsections describe them.

2.1 Unit of Phrasal Expression

The expressions appeared in the book have a coherent semantic unit as seen from examples of “in the early part of the paper” and “As a beginning, we will examine”. Namely, the expressions which do not have a coherent semantic unit such as “in

Table 1 Examples of phrasal expression

In this paper, we propose ···
To the best of our knowledge,
The rest of this paper is organized as follows.
In addition to,
With the exception of ···
with respect to ···
as we have seen
as discussed in ···
It is interesting to note that
It must be noted that

the early part of the" and "As a beginning, we will" do not considered as phrasal expressions.

Therefore, we analyzed the expressions based on a base-phrase. A base-phrase is a phrase that dominates no phrase [11] [1]. Each expression was checked if it was a sequence of base-phrases or not by using JTextPro [8] for base-phrase chunking. Consequently, out of 1,119 expressions, 1,082 (96.7%) expressions were constituted of base-phrases. Thus, we assume that a base phrase is a minimum unit of phrasal expression.

2.2 Phrasal Sign

The symbol "···" which represents abbreviation of phrases or clauses is frequently used in the book (e.g. "With the exception of ···"). 859 expressions (76.8%) appeared in the book contain the symbol. The symbol plays an important role in the using and showing of the expressions. In this research, we call the symbol *phrasal sign* and the two *phrasal signs* are used by refering to the book; <**NP**> and <**CL**>. <**NP**> and <**CL**> represent a noun phrase and a clause, respectively.

2.3 Statistical Characteristics

We found the following statistical characteristics by analyzing the book:

- **It occurs frequently.**
  The expressions appeared in the book are frequently used in academic writing.

[1] For example, the sentence "In this paper, we propose a new method." is converted into a sequence of base-phrases "[PP In] [NP this paper] , [NP we] [VP propose] [NP a new method] .". Here, parenthetical parts are base-phrases.

- **The length is not too short.**
  The expressions composed of one or two base-phrases account for only 6.9% of all expressions appeared in the book.
- **The preceding/succeeding words are various.**
  The phrasal expressions are used in various contexts. Thus, phrasal expressions can be preceded/succeeded by many kinds of base-phrases. Let us consider the expressions “in spite” (not phrasal expression) and “in spite of” (phrasal expression). As for “in spite”, term frequency was 36 and suceeding base-phrase was only “of” in the research papers used in our expriments. On the other hand, as for “in spite of”, term frequency was same as “in spite.” However, the frequency of kinds of succeeding base-phrases was 36 (e.g. “their inability”, “the noise”, “the significant error rate”, etc.). This shows that phrasal expressions have a tendency to precede/succeed various base-phrase.

2.4 Syntactic Constraints

We found such syntactic constraints as collocation between words and the availability for writing research papers. For example, “stem from” is appeared in the book. On the other hand, “stem in” and “stem with” are not appeared. Namely, collocation between words is an important factor in determining whether a given expression is phrasal expression or not. In addition, the book contains not only general expressions such as “in other words” but also specialized expressions for writing research papers such as “The purpose of this paper is to” and “The result of the experiment was that.” This shows that speciality of expressions provide a clue to identifying expressions as phrasal expression.

3 Acquisition of Phrasal Expression

Phrasal expressions are extracted from research papers based on the characteristics shown in Section 2. Figure 1 shows the processing flow for acquiring phrasal expressions. First, sequences of base-phrases are extracted from research papers. Secondly, noun phrases in them are replaced into <NP>. Note that sequences of base-phrases which contain three or more <NP> are not generated. Thirdly, sequences of base-phrases satisfying statistical constraints are acquired from them. Then, sequences of base-phrases which do not satisfy syntactical characteristics are eliminated from them. Finally, sequences of base-phrases are postfixed with <CL> if the last base-phrase of them is complementizer phrase (e.g. “that”, “which”, “so that”). The following subsections describe methods for acquiring phrasal expressions using statistical characteristics and syntactic constraints.

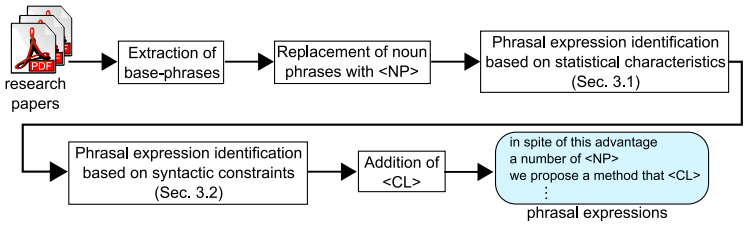

Fig. 1 Flow for acquiring phrasal expressions

3.1 Phrasal Expression Identification based on Statistical Characteristics

Candidates for phrasal expressions are extracted from sequences of base-phrases using statistical information. Note that we do not acquire sequences of base-phrases which meet conditions that relative document frequency is less than 1% or the number of base-phrases is one.

We used the scoring functions *Lscore* and *Rscore* based on Ikenos' method [1] in order to identify whether a given sequence of base-phrases has statistical characteristics. The functions are described as follows:

$$Lscore(E) = \log(tf(E)) \times length(E) \times H_l(E).$$

$$Rscore(E) = \log(tf(E)) \times length(E) \times H_r(E).$$

Here, E is a sequence of base-phrases. $length(E)$ denotes the number of base-phrases contained in E. $tf(E)$ represents term frequency of E in target research papers. $H_l(E)$ and $H_r(E)$ denote the entropy of a probability distribution of preceding and succeeding base-phrase, respectively. The scores are higher if the frequency of kinds of preceding and succeeding base-phrases are large and their frequency are uniformly. $H_l(E)$ and $H_r(E)$ are formulated by the following equations, respectively:

$$H_l(E) = -\sum_i Pl_i(E) \log Pl_i(E).$$

$$H_r(E) = -\sum_i Pr_i(E) \log Pr_i(E).$$

Pl_i/Pr_i is a probability that E is preceded/succeeded by a base-phrase Xi.

$$Pl_i(E) = P(X_iE|E) = \frac{P(X_iE)}{P(E)} \approx \frac{tf(X_iE)}{tf(E)}.$$

$$Pr_i(E) = P(EX_i|E) = \frac{P(EX_i)}{P(E)} \approx \frac{tf(EX_i)}{tf(E)}.$$

Table 2 Rules based on grammatical constraints

Pattern
A sequence does not have interrogatives, adjective phrases, noun phrases which are not only pronoun, and verb phrases which include verbs.
The first or last word of a sequence is “and”.
A sequence has complementizer phrase which are not the first and last base-phrases.
A sequence is ended with “[complementizer|;|:|,] <NP>” or “PP”.
The last word of a sequence is a nominative pronoun (“we”, “I”, “he”, “she”, “they”).
A sequence is begun with to-infinitive, a complementizer “that” or “PP <NP>”
A sequence contains to-infinitive and does not contain an infinitive verb.
A sequence contain “<NP> [of|in|,|and] <NP>” or “<NP> “(” <NP> “)””.
NP of NP
NP [of|in] <NP>the threshold of <NP>
<NP> PP NP (PP <NP>)
PP NP
PP VBG (PP is not “without”)
PP VBG (<NP>)
NP ADVP
NP interrogative
interrogative VP
pronoun (ADVP) VP (PP) (<NP>)
(NP| <NP>) copula (ADVP) (NP|ADJP) (PP)

The first, second and third terms in *Lscore* and *Rscore* represent length, term frequency and the type of preceding and succeeding base-phrases, respectively. Namely, the more the sequence reflects statistical characteristics described in Sec. 2.3, the higher the score is.

Our method considers E as candidate for phrasal expressions if E satisfies the following inequations:

$$Lscore(E) > Lscore(XE)$$

$$Rscore(E) > Rscore(EX)$$

Here, X is a preceding/succeeding base-phrase. This means that EX/XE has more base-phrase than E. If E satisfies the above two equations, E is extracted.

3.2 Phrasal Expression Identification based on Syntactic Constraints

Phrasal expressions have syntactic characteristics described in Sec. 2.4. However, since the characteristics are too various, it is difficult to identify whether a target expression has them. Therefore, in our method, sequences which do not have syntactic characteristic are eliminated by a rule-based approach.

In order to generate a rule, 809 sequences of base-phrases were extracted at random from candidates for phrasal expressions and judged whether a given sequence

Table 3 Statistics of experimental data

papers	sentences	base-phrases	words
1,232	204,788	2,683,773	5,516,612

is phrasal expression or not. We generated the rule composed of 25 patterns based on grammatical information according to the analysis of them. The generated rule is shown in Table 2. NP, VP, PP, ADVP, ADJP and VBG represent noun phrase, verb phrase, prepositional phrase, adverbial phrase, adjective phrase and gerund, respectively. Note that <NP> is different from NP. The rule is not applied if a given sequence is appeared in existing dictionaries or has specialized nouns or verbs in academic writing.

Specialized nouns and verbs are acquired by comparing relative term frequency in target research papers with the frequency in general documents such as newspapers and Web. Given a word *w*, it is identified as specialized words if it satisfies the following conditions:

- Relative document frequency in target research papers is larger than or equal to α %.
- Relative term frequency in target research papers is more than β times relative term frequency in general documents.

The thresholds α and β are set empirically.

4 Experiments

4.1 Experimental Settings

As our expermental data set, we used the proceedings of the ACL [2] from 2001 to 2008. Table 3 shows statistical information of the set. We evaluated our method which extracted 4945 phrasal expressions from experimental data. We selected Eijiro 4th Edition [9] as a dictionary used in Sec. 3.2. Specialized nouns and verbs were extracted by comparing the experimental data set with the Wall Street Journal data from the Penn Treebank [3]. The thresholds α, β for nouns and β for verbs were manually set to 1, 4 and 2, respectively, by comparing the Wall Street Journal data with the proceedings of COLING2000, COLING2002 and COLING2004. We extracted 1,119 nouns and 226 verbs with these thresholds. We used pdftext[3] to convert PDF version to plain texts and JTextPro [8] for base-phrase chunking.

As our evaluation data, 500 sequences of base-phrases were extracted from the experimental data at random and judged them by one of authors who is familiar with academic writing. We evaluated our method based on precision (the ratio of collect

[2] The Conferences of Association for Computational Linguistics

[3] http://www.foolabs.com/xpdf/

Table 4 Experimental results

	precision (%)	recall (%)	F-measure
Baseline	16.20 (81/500)	100.00 (81/81)	27.88
Statistical	23.51 (59/251)	72.84 (59/81)	35.54
Syntactic	44.07 (52/118)	64.20 (52/81)	52.26
Proposed	57.53 (42/73)	51.85 (42/81)	54.55

phrasal expressions to successfully extracted phrasal expressions) and recall (the ratio of successfully extracted phrasal expressions to all collect phrasal expressions). We compared the following four methods to evaluate our method:

Baseline phrasal expressions were acquired at random.
Statistical phrasal expressions were acquired using only statistical information.
Syntactic phrasal expressions were acquired using only syntactic information.
Proposed phrasal expressions were acquired using both statistical and syntactic information.

4.2 Experimental Result

Experimental results are shown in Table 4. Out of 500 base-phrases in the evaluation data, 81 was collect phrasal expressions. Our proposed method achieved 57.53% in precision and 51.85% in recall. In comparison with random extraction, the methods using both or either statistical and syntanctic information improved in F-measure. The results show that the use of both statistical and syntactic information is available for acquiring phrasal expressions. Therefore, we have confirmed the feasibility of our method.

Table 5 shows the examples of phrasal expression acquired successfully. Expressions appeared in dictionaries such as “As a result,” and “adding <NP> to <NP>” were acquried. Furthermore, useful expressions which are not appeared in dictionaries such as “In this paper, we propose” and “<NP> divided by the total number of <NP>” could be acquired.

4.3 Discussion

We investigated why the recall decreased by using statistical information. Out of collect 22 phrasal expressions which were not extracted, seven (31.8%) was frequently succeeded by “of” (e.g. “we have performed <NP>” and “we also show <NP> ”). They were eliminated because their *Rscore* were lower than *Rscore* for “we have performed <NP> of” and “we also show <NP> of” since noun phrase is frequently succeeded by “of” and “of” is succeeded by various noun phrases. This issue would be addressed by replacing “<NP> of <NP>” with “<NP>”.

Table 5 Examples of successfully acquired phrasal expressions

phrasal expression
<NP> is set to <NP>
<NP> is shown in Figure <digit>.
<NP> leads to <NP>,
<NP> depends on <NP>,
<NP> attached to <NP>
<NP> applied to <NP>
<NP> divided by the total number of <NP>
<NP> is not statistically significant.
<NP> is consistent with <NP>
Using <NP> as <NP>
As a result,
extracting <NP> from <NP>,
adding <NP> to <NP>.
the results obtained with <NP>
the occurrence of <NP>
N is the total number of <NP>
when <NP> are used.

We investigated why the precision did not significantly improved by using statistical information. Out of 192 incorrect phrasal expressions extracted by using statistical information, 29 (15.1%) was base-phrases whose preceding base-phrase was nominative noun phrase (e.g. "is treated as <NP>" and "is created for <NP>"). Collect phrasal expressions were "<NP> is treated as <NP>" and "<NP> is created for <NP>". However, *Lscore* for "<NP> is treated as <NP>" was not larger than *Lscore* for "is treated as <NP>" since "is treated as <NP>" was preceded by various noun phrase and "<NP> is treated as <NP>" was frequently preceded by a preposition "on". We will have to consider the formula taking nominative noun phrases into account.

5 Conclusion

In this paper, we proposed the method for acquiring phrasal expressions from research papers. The method extracts phrasal expressions from sequences of base-phrases in research papers based on statistical and syntactic information.

In this paper, phrasal expressions in the field of computational linguistics were acquired. In the future, we will apply our method to research papers in other fields although the rule based on grammatical information might have to be sophisticated. In addition, we would like to present synonymous phrasal expressions.

Acknowledgments:

This research was supported in part by the Grant-in-Aid for Scientific Research (B) (No. 20300092) of JSPS and The Hori Information Science Promotion Foundation.

References

1. Ikeno, A., Hamaguchi, Y., Yamamoto, E., Isahara, H.: Techinical term acquisition from web document collection. Transactions of Information Processing Society of Japan 47(6), 1717-1727 (2006). (in Japanese).
2. Kato, Y., Egawa, S., Matsubara, S., Inagaki, Y.: English sentence retrieval system based on dependency structure and its evaluation. In: Proceedings of 3rd International Conference on Information Digital Management, pp. 279-285 (2008).
3. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(4), 313-330 (1993).
4. Miyoshi, Y., Ochi, Y., Kanenishi, K., Okamoto, R., Yano, Y.: An illustrative-sentences search tool using phrase structure “SOUP”. In: Proceedings of 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 1193-1199 (2004).
5. Narita, M., Kurokawa, K., Utsuro, T.: A web-based English abstract writing tool using a tagged E-J parallel corpus. In: Proceedings of 3rd International Conference on Language Resources and Evaluation, pp. 2115-2119 (2002).
6. Nishimura, N., Meiseki, K., Yasumura, M.: Development and evaluation of system for automatic correction of English composition. Transactions of Information Processing Society of Japan 40(12), 4388-4395 (1999). (in Japanese).
7. Oshika, H., Sato, M., Ando, S., Yamana, H.: A translation support system using search engines. IEICE Technical Report. Data Engineering 2004(72), 585-591 (2004). (in Japanese).
8. Phan, X.H.: JTextPro: A Java-based text processing toolkit. http://jtextpro.sourceforge.net/ (2006).
9. Project, E.D.: Eijiro 4th Edition. ALC Press Inc. (2008).
10. Sakimura, K.: Useful expressions for research papers in English. Sogen-sha (1991). (in Japanese).
11. Sang, E.F.T.K., Buchholz, S.: Introduction to the CoNLL-2000 shared task: Chunking. In: Proceedings of 4th Conference on Computational Natural Language Learning and of the 2nd Learning Language in Logic Workshop, vol. cs.CL/0009008, pp. 127-132 (2000).
12. Sugino, T., Ito, F.: How to write a better English thesis. Natsume-sha (2008). (in Japanese).
13. Yamanoue, T., Minami, T., Ruxton, I., Sakurai, W.: Learning usage of english KWICLY with WebLEAP/DSR. In: Proceedings of 2nd International Conference on Information Technology and Applications (2004).