Automatic Acquisition of Useful English Expressions Using Dependency Relations

Yoshihide Kato
Information Technology Center
Nagoya University
Furo-cho, Chikusa-ku, Nagoya
464-8601 Japan

Kazuya Kuzuhara
Graduate School of Information Science
Nagoya University
Furo-cho, Chikusa-ku, Nagoya
464-8601 Japan

Shigeki Matsubara
Graduate School of Information Science
Nagoya University
Furo-cho, Chikusa-ku, Nagoya
464-8601 Japan

***Abstract*—In this paper, we propose a method to acquire English expressions useful for academic writing. The method utilizes dependency relations between words and their statistical information. It acquires subsequences of English sentences whose words are connected with dependency relations, and discriminates useful ones for academic writing by using statistical information. We conducted an experiment and the result demonstrates the effectiveness of our method.**

I. Introduction

English academic writing is a difficult task for non-native researchers. To reduce the difficulty, they often rely on lexica of English phrases (e.g. [4], [5]) to know useful expressions for English academic writing. However, the lexica do not have sufficient amounts of expressions.

To solve this problem, Kozawa et al. have proposed a method of automatically extracting phrasal English expressions from English research papers[2]. Their method extracts word sequences useful for English academic writing by using statistical information such as their frequencies. However, their method has the following drawbacks:

- The method cannot obtain the expressions which are in the form of discontinuous word sequences.
- The method may mistakenly obtain expressions whose words have no relationship.

To overcome the drawbacks, we propose a method of extracting English expressions utilizing dependency relations. Our proposed method extracts word sequences whose words are connected with dependency relations. Our method can obtain English expressions which are in the form of discontinuous word sequences, because dependency relations exist between discontinuous words. Furthermore, the obtained expressions are guaranteed to be connected with dependency relations.

To demonstrate the effectiveness of our method, we conducted an experiment. We extracted English expressions from the proceedings of the ACL from 2001 to 2008. We evaluated the precision and the recall of our method. The precision is 38.0% and the recall is 81.5%.

This paper is organized as follows: Sec. II describes what kind of expressions we want to acquire. Sec. III proposes a method of acquiring English expressions using dependency relations. Sec. IV reports an experimental result. Sec. V concludes this paper.

II. English Expressions and their Acquisition

This section explains the English expressions which we want to acquire. In this paper, the term "English expression" is used to describe a word sequence which satisfies the following conditions:

- It is useful for English academic writing.
- It can be used as an indexing term of lexica.

As an example, let us consider the following word sequences:

1) In this paper, we describe ...
2) The reason why ... is that ...
3) For instance, it ...

Word sequences 1) and 2) are English expressions. These word sequences are useful for English academic writing. On the other hand, word sequence 3) is not an English expression. Though it may be useful for English academic writing, we do not want to use it as an indexing term. Because the word "it" is unnecessary. The word sequence "For instance, ..." is more appropriate for the indexing term. Our aim is to automatically acquire word sequences such as 1) and 2) and to discard word sequences such as 3).

A. Previous Works

Kozawa et al. proposed a method of extracting English expressions from research papers[2]. Their method discriminates English expressions using statistical information about word sequences. The method cannot obtain discontinuous English expressions such as 2). The reason is that the method first extracts continuous word sequences which occur in research papers and selects English expressions from the word sequences. Because the word sequence "The reason why" and the word sequence "is that" in English expression 2) usually appear separately in sentences, the word sequence "the reason why ... is that ..." is not extracted at the first stage. Furthermore, the method may mistakenly obtain word sequence 3) as an English expression, when it frequently occurs in the research papers.

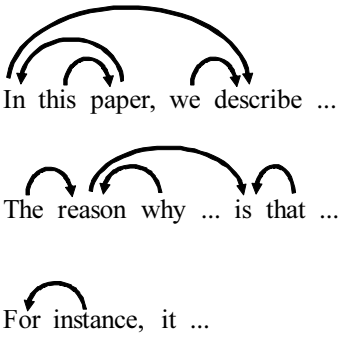

Fig. 1. dependency relations in English expressions

B. Syntactic Characteristic of English Expressions

To solve the problem described in the previous section, we focus on dependency relations which appear in English expressions. Dependency relations refer to certain kinds of relations between words, such as modifier-modifee relations, predicate-argument relations and so on. From the viewpoint of dependency relations, word sequence 1) and 2) are connected with dependency relations (see Fig.1). On the other hand, word sequence 3) is not connected with dependency relations. In general, English expressions are connected with dependency relations. An important point is that discontinuous English expression 2) is connected with dependency relations. By extracting word sequences whose words are connected with dependency relations, we expect that discontinuous English expressions such as 2) can be obtained and that the number of the word sequences such as 3) which are not English expressions is reduced.

III. English Expression Acquisition Based on Dependency Relations

This section proposes a method of extracting English expressions using dependency relations. The outline of our proposed method is as follows:

1) English sentences in research papers are parsed by a dependency parser. Each parsed sentence is represented as a dependency tree.
2) Frequent tree patterns are extracted from the dependency trees. The tree patterns corresponds to continuous or discontinuous word sequences.
3) Tree patterns which correspond to English expressions are discriminated using statistical information about dependency relations.
4) Word sequences are recovered from the discriminated tree patterns.

A. Tree Representation of Dependency Relations

Our proposed method focuses on dependency relations. Dependency relations refer to certain kinds of relations between words. In a dependency relation, one word is called *head*, and the other word is called *dependent*. When a word w_d is a dependent of a word w_h, we say that w_d depends on w_h, and write $w_d \rightarrow w_h$. The dependency relations in a sentence can be represented as a tree structure. Each node is labeled with a word in the sentence. When a word w_d depends on a word w_h, the node labeled with w_d is a child of the node labeled with w_h. By extracting tree patterns which occur in the dependency trees, our method obtains word sequences whose words are connected with dependency relations. However, word sequences cannot be recovered from the tree patterns, because they lack the following information:

1) There is no information about word order. When a node labeled with w_d is a child of a node labeled with w_h, we cannot determine whether w_d precedes w_h or not.
2) There is no information about abbreviations of constituents, which are represented by the symbol "..." in English expressions.

To avoid this problem, we extend the tree structures as follows:

1) Each node which corresponds to a word has two children. The children's labels are LEFT and RIGHT. If a word w_d depends on a word w_h from left, the node w_d is the child of the LEFT node of the node w_h. If w_d depends on w_h from right, the node w_d is the child of the RIGHT node of the node w_h. By using this additional nodes, we can determine the order of the words in a dependency tree.
2) Each node has additional information about constituents. For example, if the words dominated by a node constitute a noun phrase, the node is labeled with $\langle$NP$\rangle$.

B. Extracting Tree Patterns

To obtain word sequences whose words are connected with dependency relations, our method parses sentences in English research papers by using a dependency parser, and extracts tree patterns from the parsed sentences. Because it is intractable to extract all tree patterns, we uses a tree mining algorithm based on FREQT[1]. FREQT extracts frequent tree patterns from a set of trees. At first, FREQT extracts frequent tree patterns with size 1, which consist of single nodes. Then, FREQT iteratively enumerates candidate tree patterns with size k by attaching new nodes to the frequent tree patterns with size $k-1$, and discards the tree patterns whose frequencies is less than a threshold.

We extend FREQT to deal with the symbol "..." in English expressions. In our method, several nodes in the dependency trees have labels which represent constituents. The labels correspond to the symbol "...". For example, The English expression "The reason why ... is that ..." are obtained from the tree pattern shown in Fig. 2. To extract this kind of tree patterns, we extend the enumeration of tree patterns as follows:

- In attaching node phase, if the attached node has constituent label, our method creates two tree patterns: the one is the result of attaching a node labeled with its word label, the other is the result of attaching a node labeled with its constituent label.

Furthermore, our method does not attach a node in the following case:

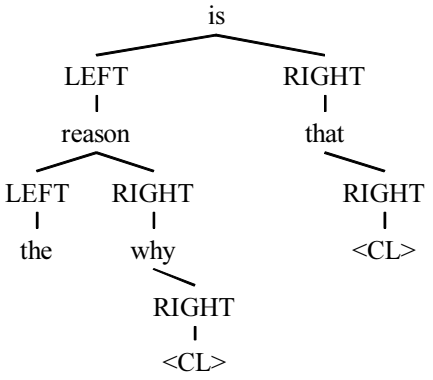

Fig. 2. the tree pattern corresponding to the expression "the reason why ... is that ..."

- New node is not attached to the node with a constituent label of a tree pattern.

C. Discriminating English Expressions

Previous section describes a method of extracting frequent tree patterns. By recovering word sequences from the tree patterns, we can obtain frequent word sequences whose words are connected with dependency relations. Here, note that not all the frequent word sequences are English expressions. Our method selects English expressions from the frequent word sequences by using statistical information about tree patterns.

1) Discrimination Based on Statistical Information: First of all, let us consider the following word sequences:

1) the reason why ... is
2) the reason why ... is that

Word sequence 2) includes word sequence 1). If word sequence 1) frequently co-occurs the word "that", word sequence 2) is more useful than 1). That is, we want to accept word sequence 2) and discard word sequence 1).

Our method discards word sequences which are included by useful word sequences by using statistical information. In order to do so, for each node in the tree patterns, our method assesses whether a specific word or constituent is attached to it. We define an entropy of a node label. That the entropy of a node label of a tree pattern is low means that the node frequently labeled with a specific word or constituent. That is, the possibility that the tree pattern is a part of an English expression is high. Our method discards the tree patterns.

In the following, we define the entropy of a node label of a tree pattern. First of all, we introduce some notations. Let D be a set of dependency trees, which are obtained from English research papers. Let T and n be a tree pattern and a node in T, respectively. $Trees(T)$ is a set of trees which is defined as follows:

$$Trees(T) = \{T_D \mid T_D \text{ matches } T \text{ and } T_D \text{ occurs in } D\}$$

$Trees(T)$ is a set of trees which match T. $Nodes(n, T)$ is a set of nodes which is defined as follows:

$$\begin{aligned} &Nodes(n, T) \\ &= \{m \mid \exists T_D \in Trees(T)[m \text{ is a node of } T_D \\ &\qquad \text{and } m \text{ corresponds to } n]\} \end{aligned}$$

$Nodes(n, T)$ is a set of nodes which corresponds to n. For a node n of a tree pattern T, we define four types of adjacent nodes of n as follows:

$$\begin{aligned} &A_{dr}(n, T) \\ &= \{m'' \mid \exists m, m'[m \in Nodes(n, T) \\ &\qquad \text{and } m' \text{ is the parent of } m \\ &\qquad \text{and } m' \text{ has the label } RIGHT \\ &\qquad \text{and } m'' \text{ is the parent of } m']\} \end{aligned}$$

$$\begin{aligned} &A_{dl}(n, T) \\ &= \{m'' \mid \exists m, m'[m \in Nodes(n, T) \\ &\qquad \text{and } m' \text{ is the parent of } m \\ &\qquad \text{and } m' \text{ has the label } LEFT \\ &\qquad \text{and } m'' \text{ is the parent of } m']\} \end{aligned}$$

$$\begin{aligned} &A_{hr}(n, T) \\ &= \{m'' \mid \exists m, m'[m \in Nodes(n, T) \\ &\qquad \text{and } m' \text{ is a child of } m \\ &\qquad \text{and } m' \text{ has the label } RIGHT \\ &\qquad \text{and } m'' \text{ is a child of } m']\} \end{aligned}$$

$$\begin{aligned} &A_{hl}(n, T) \\ &= \{m'' \mid \exists m, m'[m \in Nodes(n, T) \\ &\qquad \text{and } m' \text{ is a child of } m \\ &\qquad \text{and } m' \text{ has the label } LEFT \\ &\qquad \text{and } m'' \text{ is a child of } m']\} \end{aligned}$$

We define $EX(T, n, type)$ as a set of labels with which the adjacent nodes of n are labeled as follows (see Fig.3):

$$\begin{aligned} &EX(T, n, type) \\ &= \{l \mid \exists m \in A_{type}(n, T) \\ &\qquad [l \text{ is the label of } m \\ &\qquad\quad \text{and } m \text{ does not correspond to any node in } T] \\ &\qquad \text{or } [A_{type}(n, T) = \emptyset \text{ and } l = null]\} \end{aligned}$$

$null$ is a special symbol which represents that n has no adjacent nodes. We define an entropy of labels of adjacent nodes as follows:

$$\begin{aligned} &H_{type}(L \mid T, n) \\ &= -\sum_{l \in EX(T,n,type)} P_{type}(L = l \mid T, n) \log P_{type}(L = l \mid T, n) \end{aligned}$$

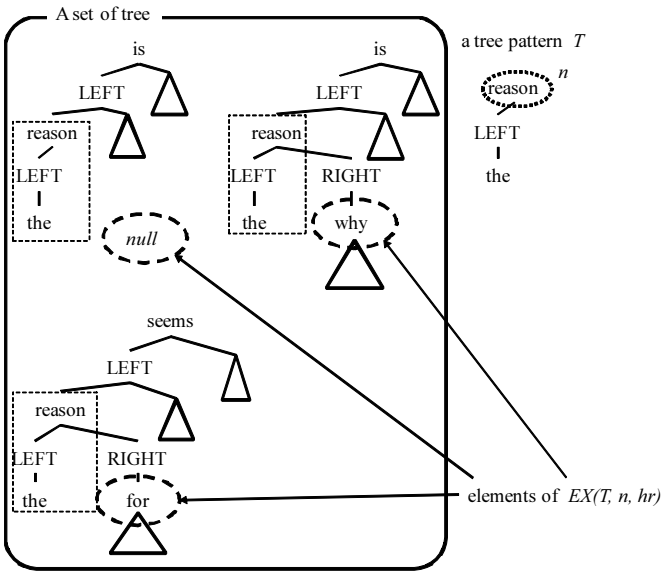

Fig. 3. an example of $EX(T, n, type)$

where L is a random variable which represents node labels. $P_{type}(L = l \mid T, n)$ is the probability that the label of an adjacent node of n is l. The probability is defined as follows:

$$P_{type}(l \mid T, n) = \frac{C(expand(T, n, type, l))}{\sum_{l \in EX(T,n,type)} C(expand(T, n, type, l))}$$

where $C(\cdot)$ is the frequency of a tree pattern. $expand(T, n, type, l)$ is the tree pattern which is obtained by attaching a node with label l. When $type = dr$, the procedure is as follows:

1) A node labeled with $RIGHT$ is attached to n as its parent (Let n' be the attached node.).
2) A node labeled with l is attached to n' as its parent.

Our method utilizes the entropy measure to discriminate word sequences which are parts of English expressions. Our method discards a tree pattern T, if the following conditions hold for some n in T and some $type \in \{dr, dl, hr, hl\}$:

1) $P_{type}(null|T, n) < \alpha$
2) $H_{type}(L|T, n) < \beta$

α and β are thresholds. Condition 1) means that the node n frequently has adjacent nodes. Condition 2) means that the adjacent nodes of n has a specific label.

IV. An Experiment

To demonstrate the effectiveness of our method, we conducted an experiment. We used 165,116 sentences in the proceedings of the ACL from 2001 to 2008, and 500 word sequences as evaluation data, which are extracted from the sentences. For each word sequence, whether it is an English expression or not is manually judged. The sentences are parsed using the parser Enju[3] and converted to dependency relations using Pennconverter[6].

TABLE I
AN EXPERIMENTAL RESULT

	precision(%)	recall(%)	F-value
our method	37.8%	81.5%	51.7
Kozawa et al.	23.5%	72.8%	35.5

We set α and β to 0.5 and 1.3, respectively. We extracted 1,925,449 tree patterns from the proceedings. Our method selects 127,059 word sequences obtained from tree patterns as English expressions. We measured the precision and the recall of our method using the evaluation data. The result is shown in Table I. Our method outperforms the previous method based on word sequences. This result demonstrates the effectiveness of our method based on dependency relations.

We show examples of English expressions obtained by our method.

- the fact that ... suggests that ...
- since ... we can conclude that ...

These examples show that our method can acquire English expressions which are in the form of discontinuous word sequences.

V. Conclusion

This paper proposed a method of acquiring English expressions using dependency relations. Our method represents sentences as dependency trees, and extracts frequent tree patterns to obtain frequent word sequences whose words are connected with dependency relations. Our method selects English expressions for academic writing from the obtained word sequences using statistical information about dependency relations. An experimental result demonstrates that the dependency relation-based approach outperforms the word sequence based approach.

Our method cannot obtain less frequent English expressions. To overcome this problem, we will use a huge amount of research papers to acquire English expressions.

Acknowledgment

This research is partially supported by the Kayamori Foundation of Informational Science Advancement.

References

[1] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto and S. Arikawa: Efficient Substructure Discovery from Large Semi-structured Data, *Proc. of 2nd SIAM Int. Conf. on Data Mining*, pp.158–174, 2002.
[2] S. Kozawa, Y. Sakai, K. Sugiki and S. Matsubara: Automatic Extraction of Phrasal Expressions for Supporting English Academic Writing, *Proc. of 2nd Int. Symposium on Intelligent Decision Technologies*, pp.485–493, 2010.
[3] Y. Miyao and J. Tsujii: Feature Forest Models for Probabilistic HPSG parsing, *Computational Linguistics*, 34(1), pp.35–80, 2008.
[4] K. Sakimura: Useful Expressions for Research papers in English, Sogensha (1991) (in Japanese)
[5] T. Sugino and F. Ito: How to Write a Better English Thesis, Natsume-sha (2008) (in Japanese)
[6] http://fileadmin.cs.lth.se/nlp/software/pennconverter/pennconverter.jar