Searching for Illustrative Sentences for Multiword Expressions in a Research Paper Database

Hidetsugu Nanba

Hiroshima City University, 3-4-1 Ozuka-higashi, Asaminami-ku
Hiroshima, 731-3194, Japan
nanba@its.hiroshima-cu.ac.jp

Satoshi Morishita

NEC Micro Systems, 1-403-53,
Kosugicho, Nakaharaku, Kawasaki
211-0063, Japan
morishita@nlp.its.hiroshima-cu.ac.jp

Abstract. We propose a method to search for illustrative sentences for English multiword expressions (MWEs) from a research paper database. We focus on syntactically flexible expressions such as "regard – as." Traditionally, illustrative sentences that contain such expressions have been searched for by limiting the maximum number of words between the component words of the MWE. However, this method could not collect enough illustrative sentences in which clauses are inserted between component words of MWEs. We therefore devised a measure that calculates the distance between component words of an MWE in a parse tree, and use it for flexible expression search. We conducted experiments, and obtained a precision of 0.832 and a recall of 0.911.

Keywords: multiword expressions, a support system for writing technical documents, illustrative sentence, a research paper database

1 Introduction

When non-English native speakers write or translate technical documents using English, they are often confused about how to choose proper expressions. Illustrative sentences shown with each entry word in dictionaries are useful for selecting the most appropriate expression from candidates. However, these sentences are not always useful when non-native speakers write technical documents, because while some expressions that are commonly used but not in a specific research domain are included in dictionaries, some technical expressions that are commonly used in the specific domain are not usually included in dictionaries. Therefore, a support system for writing technical documents is required. In this paper, we propose a method for searching for illustrative sentences of English multiword expressions (MWEs) from a set of research papers in a specific domain.

Nanba et al.[7,8] constructed a multilingual research paper database, "PRESRI", by collecting more than 78,000 Postscript and PDF files published on the Internet. The database contains research papers in domains such as computer science, nuclear biophysics, chemistry, astronomy, material science and electrical engineering.

To collect research papers in a specific domain from PRESRI, we can use keyword search and citation analysis, such as bibliographic coupling [5] and co-citation analysis [10]. As PRESRI possesses information about the sources (journal titles or conference names) of research papers, we can collect illustrative sentences of MWEs that were commonly used in particular conferences or journals. We construct a system that searches for illustrative sentences of English MWEs from research papers from the PRESRI collection.

The remainder of this paper is organized as follows. In Section 2, we describe multiword expressions. In Section 3, we explain our method for searching for illustrative sentences of a given MWE. To investigate the effectiveness of our method, we conducted some tests. In Section 4, we report the results, and conclude in Section 5.

2 Multiword Expressions

Expressions that consist of multiple words are called Multiword Expressions (MWEs). Baldwin [3] classified MWEs as follows.

1. **Lexicalized phrases**
   1. Fixed expressions
      Fixed strings that undergo neither morphosyntactic conversion nor internal modification (e.g., *ad hoc*).
   2. Semi-fixed expressions
      Expressions that adhere to strict constraints on word order and composition, but undergo some lexical variation. For example, the word "oneself" in an MWE "prostrate oneself" has some variations, such as "himself" or "herself." Compound nouns are also included in this category.
   3. Syntactically flexible expression
      Expressions in which more than one word are inserted between their component words (e.g., *take the evidence way*).
2. **Institutionalized phrases**
   In terms of syntax and semantics, these are considered as MWEs (e.g., *kindle excitement*).

In our work, we focus on searching for illustrative sentences of syntactically flexible expressions, because fixed expressions and semi-fixed expressions are easy to search for by conducting simple string matching after stemming words in target sentences. On the other hand, some restrictions are necessary when searching for illustrative sentences containing flexible expressions. In the next section, we will explain our method of searching for such sentences.

3 Searching for Illustrative Sentences of Flexible expressions

3.1 Related Works

Verb–particle construction (VPC) is a kind of "syntactically flexible expression" that consists of a verb and a particle, such as "hand in". Baldwin [1,2] proposed the following methods to extract VPCs from texts.

1. Extract VPCs if the number of words between a particle and its governing verb is less than five.
2. Extract VPCs using method one with the restriction that the inserted words are nouns, prepositions, or verb chunks.
3. Extract VPCs using method two with a chunk grammar.

The limitation of "less than five words" has also been used for extracting collocations [9]. However, the limitation of "less than five words" does not ensure that we can search for illustrative sentences for MWEs other than VPCs comprehensively, because it is not uncommon that long phrases or clauses are inserted between component words of MWEs other than VPCs. To show the variety of illustrative sentences, we search for sentences in which more than four words are inserted between component words.

3.2 Our Method

Following is an illustrative sentence for the MWE "share – with."

> But Mr. Foley predicted few economic policy changes ahead, commenting that Mr. Major **shares** a very similar view of the world **with** Mr. Lawson.

The traditional method cannot detect this sentence, as there are seven words between the component words of the MWE. In Figure 1, we show a syntactic tree of this sentence. From this figure, we can find that "share" and "with" are close to each other on the tree. We therefore focus on syntactic trees for the detection of illustrative sentences.

Here, we define a measure for calculating distance between words on a syntactic tree. Figure 2 shows a flexible expression that is constructed from two words. CW and OW indicate component words of an MWE and other words, respectively. A hierarchical distance is defined as the number of nodes on the shortest path from one component word to another. In this example, as there are three nodes on the shortest path, which is shown as a bold line, the hierarchical distance is three. We extract all sentences with a hierarchical distance between components of an MWE that is smaller than a threshold value. Together with the hierarchical distance, we also use the following two definitions: "restriction of changing voice" and "insertion of a phrase".

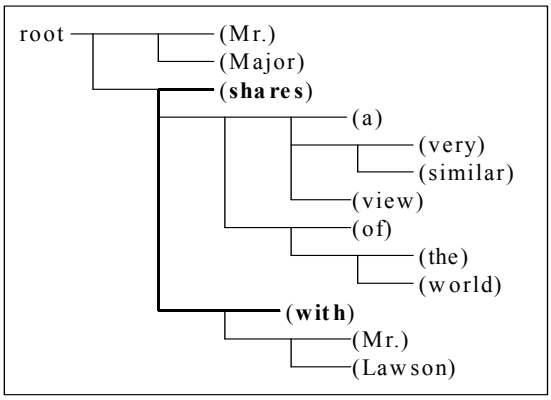

Fig. 1. A syntactic tree of an illustrative sentence for the MWE "share - with"

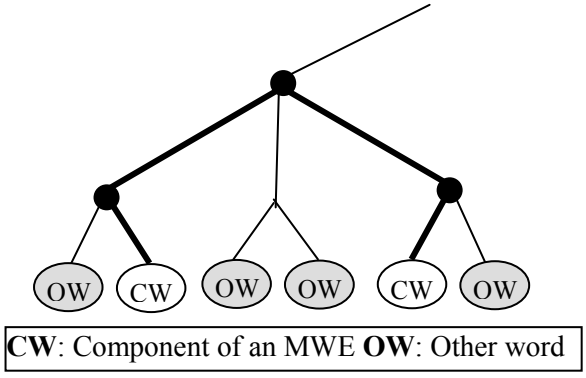

Fig. 2. An example of a hierarchical distance between component words.

Restriction of Changing Voice

When an MWE contains a transitive verb, it is possible to change voice. However, it is considered that changing voice is not a general usage. If an MWE is generally used in the passive voice, such as "be attributed to", the entry in dictionaries is also written in passive voice.

To confirm the validity of this assumption, we selected 21 MWEs that contain transitive verbs and investigated whether the voices of the MWEs are the same as those in illustrative sentences in three dictionaries: "Collins CoBUILD", "Readers Plus" (Kenkyusha, Ltd.), and "New College English Japanese Dictionary" (Kenkyusha, Ltd.). The results are shown in Table 1.

Table 1. The ratio of illustrative sentences with voices that are different from those of entries in dictionaries

Dictionary	Ratio
Collins CoBUILD	0.11 (2/17)
Readers Plus (Kenkyusha)	0.13 (2/15)
New College English-Japanese Dictionary (Kenkyusha)	0.15 (2/13)
Total	**0.13 (6/45)**

There were 45 illustrative sentences for the 21 MWEs in the three dictionaries, and the voices of MWEs differ from those in the illustrative sentences in six cases (13%). Because the number of illustrative sentences used in this investigation was small, we cannot derive a concrete conclusion. However, the results do indicate that changing voice is not a general usage. Therefore, we do not search for sentences with voices that are different from MWEs.

Restriction of Insertion of Clauses between Component Words of MWEs

There are cases when clauses are inserted between component words of MWEs. To search for such illustrative sentences, we use the following restrictions. When a clause begins between the component words of an MWE and does not end in the same split part, we consider that the clause is not a parenthetic clause, and eliminate the sentence from the candidates of illustrative sentences. Figure 3 shows an example in which a clause, shown as a shadowed rectangle does not end in the split part. In this case, we consider that this is not an illustrative sentence for an MWE "the same A as B".

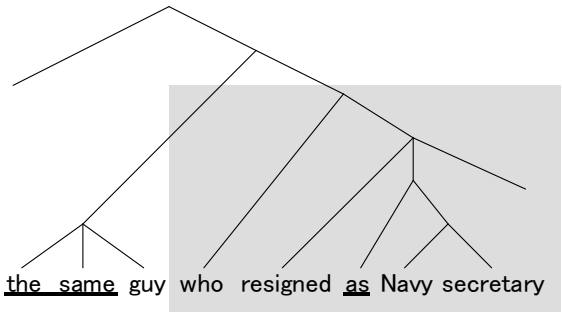

Fig. 3. An example in which a clause is not terminated between component words of an MWE

4 Experiments

To investigate the effectiveness of our method, we conducted some experiments.

4.1 Evaluation

Experimental Method

We used a syntactic parser [4] to search for illustrative sentences. The performance of the parser affects the search results directly; however, we could not estimate its effect.

Therefore, we tested our method in two ways: (1) using manually annotated syntactic tags and (2) using the results from the syntactic parser and confirming the effects of parse errors by comparing their results.

To confirm the effects of parse errors, we used Penn Treebank (PTB)[1] [6]. PTB is a large corpus of *Wall Street Journal* material, in which 74,000 sentences are manually annotated with part-of-speech tags and syntactic tags. First, we tested our methods using PTB with manually annotated syntactic tags. Second, we tested using PTB with the results from the syntactic parser. Finally, we tested using 18,000,000 sentences in PRESRI with the results from the syntactic parser.

Alternatives

We conducted tests using the following five methods.

Our methods:

(A) Using a hierarchical distance. The maximum distance was four.
(B) (A) + restriction of changing voice.
(C) (B) + restriction of insertion of a clause.

Baseline methods:

(i) String matching. The number of words in a split area was not limited.
(ii) String matching. The maximum number of words in a split area was three.

Here, we experimentally determined the threshold value as four in method A, using the data for making rules that we will describe later. In the same way, the threshold value for baseline method ii was determined as three.

Test Collections

We manually selected 53 flexible expressions from nine books about technical writing for Japanese. We use 42 MWEs for making rules and 11 for evaluation.

We constructed test collections using the following three steps:

1. Convert all words in MWEs and in all sentences into their original forms using LimaTK [11];
2. Collect all sentences using simple pattern matching;
3. Manually identify whether the sentences collected in Step 2 are valid illustrative sentences for the given MWEs.

Table 2 shows the data that we used in our examinations.

[1] http://www.cis.upenn.edu/~treebank/

Table 2. Data for the examinations

		The number of MWEs	The number of sentences for search	The number of correct sentences
PTB	For making rules	42	662	429
	For evaluation	42	351	219
PRESRI		53	2466	1720

Evaluation Measures

We evaluate our methods and baseline methods using the following equations.

$$Precision = \frac{\begin{array}{l}The\ number\ of\ sentences\ \ that\\ a\ system\ detected\ correctly\end{array}}{\begin{array}{l}The\ number\ of\ sentences\\ that\ a\ system\ detected\end{array}} \quad (1)$$

$$Recall = \frac{\begin{array}{l}The\ number\ of\ sentences\ \ that\\ a\ system\ detected\ correctly\end{array}}{\begin{array}{l}The\ number\ of\ sentences\\ that\ should\ be\ detected\end{array}} \quad (2)$$

4.2 Results

In Table 3, we show the experimental result using the data of PTB with manual parse trees. As Table 3 shows, our methods are superior to both baseline methods.

Table 3. Results of searching for illustrative sentences using PTB (Manual)

		Precision	Recall
Baseline methods	i	0.624 (219/351)	1.000 (219/219)
	ii	0.708 (155/219)	0.708 (155/219)
Our methods	A	0.796 (207/260)	0.945 (207/219)
	B	**0.868** (204/235)	0.932 (204/219)
	C	**0.868** (203/234)	0.927 (203/219)

We also show the results using the data of PTB with parse trees by the parser. The results using the statistical parser (Table 4) are better than those using manual parse trees (Table 3), because most of the sentences that could not be analyzed by the parser happened to be incorrect as illustrative sentences.

Table 4. Results of searching illustrative sentences using PTB (Parsing)

		Precision	Recall
Baseline methods	i	0.624 (219/351)	1.000 (219/219)
	ii	0.708 (155/219)	0.708 (155/219)
Our methods	A	0.880 (205/233)	0.936 (205/219)
	B	0.887 (204/230)	0.932 (204/219)
	C	**0.889** (201/226)	0.918 (201/219)

Table 5. Results of searching illustrative sentences using PPRESRI (Parsing)

		Precision	Recall
Baseline methods	i	0.697 (1720/2466)	1.000 (1720/1720)
	ii	**0.870** (1140/1311)	0.663 (1140/1720)
Our methods	A	0.841 (1303/1549)	0.758 (1303/1720)
	B	0.849 (1277/1505)	0.742 (1277/1720)
	C	0.862 (1248/1447)	0.726 (1248/1720)

Finally, we show the result using the data of PRESRI with parse trees by the statistical parser. Baseline methods ii is superior to others, while the recall of this method is the worst.

4.3 Discussions

Comparison of Baseline Method ii and our Methods

The gap of precision between method ii and our methods is more than 0.1 in tests using PTB data (Tables 3 and 4), while the gap was almost the same in the test using PRESRI data (Table 5). This is caused by the low performance of the syntactic parser with the PRESRI data. As the syntactic parser was trained using PTB, we cannot obtain the same performance for PRESRI as for PTB.

Effectiveness of Our Methods

Among our three methods, the precision of method C is the best. However, the precision of baseline method ii is superior to method C, although recall is the worst because the method eliminated all illustrative sentences if more than four words were inserted between component words of MWEs. However, high recall is also required in terms of variety of illustrative sentences.

Combination of Method C and Baseline Method ii

Method C can find illustrative sentences correctly, even when many words are inserted between the component words of the MWEs, while baseline method ii can also find sentences correctly when less than four words are inserted between component words of MWEs. Therefore, it is considered that these methods can find many different sentences, i.e. it is possible to improve recall by combining both methods.

We investigated the relations between recall and precision and threshold values of method C and baseline method ii. We show the results in Figure 4. The figure shows

that the precision of baseline method ii decreases when the threshold value exceeds three. On the other hand, the precision of method C is the highest when the threshold value of the hierarchical distance is four, then decreases as the threshold value increases.

We therefore combine baseline methods ii and C. When the threshold value of baseline method ii is smaller than a value *n*, we applied the baseline method, and when the value is larger than n, we applied method C. As a threshold value for a hierarchical distance, we used four.

We searched illustrative sentences using the combination method while changing the value of *n* from two to five. The results are shown as triangles in Figure 4. The figure shows that the combination method can improve recall while maintaining precision. When $n = 4$, we obtained the precision of 0.832 and recall of 0.911. From this result, we can conclude that the simple string matching method is useful when less than five words are inserted between component words and that using a hierarchical distance is also useful when more than four words are inserted between components.

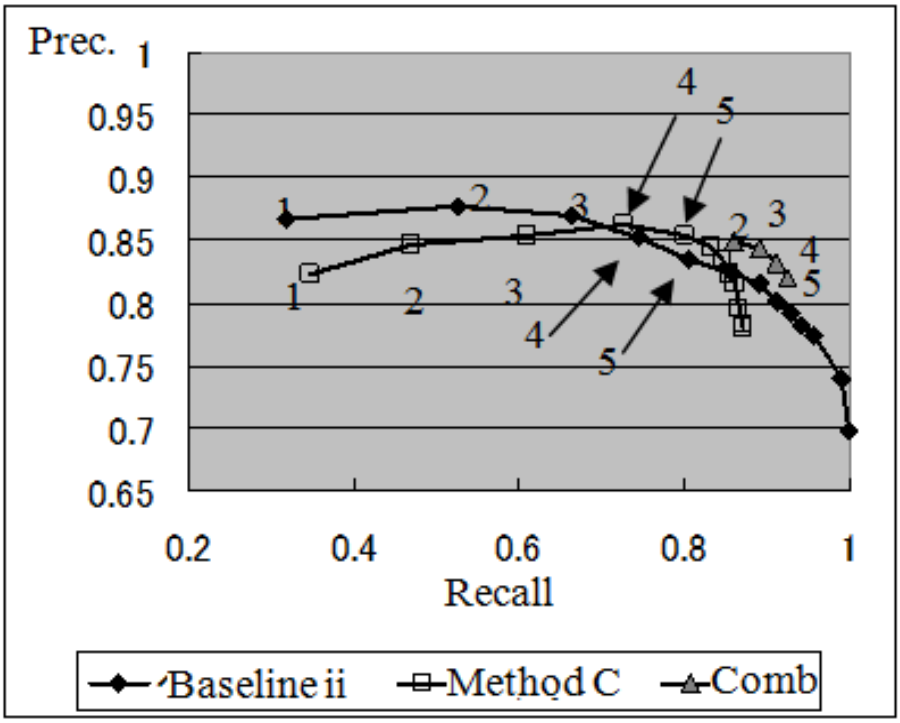

Fig. 4. Recall and Precision by baseline method ii, method C, and their combination method

5. Conclusions

We have proposed a method to search illustrative sentences of flexible expressions from the research paper database PRESRI. We conducted tests, and obtained the precision of 0.832 and recall of 0.911. From the results of the experiments, we can conclude that the simple string-matching method is useful when less than five words are inserted between component words, and that using a hierarchical distance is also useful when more than four words are inserted between components.

References

1. Baldwin, T., Villavicencio, A.: Extracting the Unextractable: A Case Study on Verb-particles. In Proceedings of the 6th Conference on Natural Language Learning 2002, pp. 98-104. (2002)
2. Baldwin. T., Bannard, C., Tanaka, T., Widdows, D.: An Empirical Model of Multiword Expression Decomposability. In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pp.89-96 (2003)
3. Baldwin, T.: Multiword Expressions. Advanced course at the Australasian Language Technology Summer School (2004)
4. Bikel, D.M.: A Distributional Analysis of a Lexicalized Statistical Parsing Model. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, a Meeting of SIGDAT, pp.182-189. (2004)
5. Kessler, M.M.: Bibliographic Coupling between Scientific Papers. In Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, pp.68-81. (1997)
6. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The Penn Treebank: Annotating Predicate Argument Structure. In Proceedings of the Human Language Technology Workshop, pp.114-119. (1994)
7. Nanba, H., Abekawa, T., Okumura, M., Saito, S.: Bilingual PRESRI: Integration of Multiple Research Paper Databases. In Proceedings of the RIAO 2004, pp.195-211. (2004)
8. Nanba, H. Kando, N., Okumura, M.: Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation. In Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop, Classification for User Support and Learning, pp.117-134. (2000)
9. Smadja, F.: Retrieving Collocations from Text: Xtract. Computational Linguistics, Vol.19, Issue 1, Special Issue on Using Large Corpora, pp.143-177. (1993)
10. Small, H.: Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents. Journal of the American Society for Information Science, Vol.24, pp.265-269. (1973)
11. Yamashita, T., Matsumoto, Y.: Language Independent Morphological Analysis. In Proceedings of the 6th Conference on Applied Natural Language Processing, pp.232-238. (2000)