Identification of Flexible Multiword Expressions with the Help of Dependency Structure Annotation

Ayaka Morimoto, Akifumi Yoshimoto, Akihiko Kato,
Hiroyuki Shindo, and Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
{morimoto.ayaka.lw1, akifumi-y, kato.akihiko.ju6,
shindo, matsu}@is.naist.jp

Abstract

This paper presents our ongoing work on compilation of English multi-word expression (MWE) lexicon and corpus annotation. We are especially interested in collecting flexible MWEs, in which some other constituents can intervene the expression such as "a number of" vs "a large number of" where a modifier of "number" can be placed in the expression while inheriting the original meaning. We first collect possible candidates of flexible English MWEs from the web, and annotate all of their occurrences in the Wall Street Journal portion of OntoNotes corpus. We make use of word dependency structure information of the sentences converted from the phrase structure annotation. This process enables semi-automatic annotation of MWEs in the corpus and simultaneously produces the internal and external dependency representation of flexible MWEs.

1 Introduction

Multiword Expressions (MWEs) are roughly defined as those that have "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag, 2002), and are classified into the following categories:

- Lexicalized phrases
  - fixed expressions: Those with fixed word order and forms (e.g. *with respect to*).
  - semi-fixed expressions: Those with lexical variation such as inflection, etc. (e.g. *keep up with, kept up with*).
  - syntactically flexible expressions: Those with a wide range of syntactic variability (e.g. some of the internal words in an MWE can have a modifier).
- Institutionalized phrases
  - Phrases that are syntactically compositional but semantically specific (e.g. *traffic light*).

In this paper we mainly focus on English syntactically flexible multi-word expressions, since they are less investigated than other types of MWEs. There are a number of MWEs that grammatically behave as single lexical items belonging to some specific parts-of-speech, such as adverbs, determiners, prepositions, subordinate conjunctions, and so on. Other than MWEs with those functions, we also consider multi-word verbs such as *take into consideration*, but not multi-word nouns. The reason we do not consider multi-word nouns is that most of them are syntactically not flexible, meaning they do not allow to have modifiers within them. MWEs have specific grammatical functionalities and can be regarded as an important part of an extended lexicon.

The objective of our work is to construct a wide coverage English syntactically flexible MWE lexicon, to describe their structures in dependency structures with possible modifiers within them, and to annotate their occurrences in the Wall Street Journal portion of OntoNotes corpus (Pradhan et al., 2007).

There have been some attempts for constructing English MWE lexicon. An English fixed MWE lexicon and a list of phrasal verbs are presented in (Shigeto et al., 2015) and (Komai et al., 2015). They

Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces,
pages 102–109, Osaka, Japan, December 11 2016.

also annotated the occurrences of those expressions in Penn Treebank. While most of English dictionaries for human use include a large list of multi-word expressions and idioms, to the best of our knowledge, there has been no comprehensive lexicon of English flexible MWEs constructed that is usable for NLP tasks.

The main contributions of our work are the following:

1. We constructed a large scale syntactically flexible English multi-word lexicon by collecting them from various web sites that list MWEs.

2. Through annotation of collected MWEs in OntoNotes corpus, we identified possible modifiers that can appear within the expressions.

Our current work has clear limitation that we cannot know how flexible those expressions are that do not appear in a form of flexible usage in OntoNotes. So, the first contribution is still ongoing. But, for the second contribution, we try to annotate all the occurrences of flexible MWEs in OntoNotes. In the following sections, we first describe related research, then explain how we collected English flexible MEWs and the method we used to annotate the occurrences of MWEs in OntoNote. We also give some statistics concerning with our experiments.

2 Related Works

Corpus annotation of MWEs hasn't been done in large scale in English. On the other hand, in French there is a large scale MWE annotated corpus (Abeillé et al., 2003), which includes 18,000 sentences annotated with 30,000 MWEs. In English, (Schneider et al., 2015) constructed an MWE-annotated corpus on English Social Web Corpus with all types pf English MWEs. However, the size of the corpus is small (3,800 sentences). This is the first and only corpus that has annotation of syntactically flexible English MWEs.

For English fixed and semi-flexible MWEs there are some works on construction of lexicons and on annotation on a large scale corpus. (Shigeto et al., 2015) and (Kato et al., 2016) annotated the Wall Street Journal portion of OnteNotes with fixed functional MWEs. The size of the corpus is 37,000 sentences and the number of annotated MWEs is 6,900. In the former work they constructed an English fixed MWE lexicon and annotated the spans of all occurrences of MWEs in the corpus. The latter work annotated and modified dependency structure of the sentences in accordance with their functionality. A specific type of English MWEs, phrasal verbs, are annotated on the same corpus by (Komai et al., 2015), in which they annotated 22,600 occurrences of phrasal verbs in 37,000 sentences.

PARSEME (PARSing and Multi-word Expressions) Project[1] is a project devoted to the issue of Multiword Expressions in parsing and in linguistic resources in multi-lingual perspective. A comprehensive introduction of the project is found in (Savary et al., 2015). A detailed survey of MWEs in Treebanks is found in (Rosén et al., 2016).

3 Collection of English Flexible Multi-word Expressions

For collecting English flexible MWE candidates, we explored ALL IN ONE English learning site[2] and the index of the English Idiom dictionary by Weblio[3]. Both sites provide useful information for English learners such as dictionaries, examples and useful expressions. In addition, we explored the entiries in Wiktionary[4] that contain white space(s) within the expressions whose part-of-speech are either Verb, Adjective, Adverb, Preposition, or Conjunction.

All of those collected 16,339 MWE candidates. Then, we counted the corpus occurrences of those expressions using Web 1T 5-gram(LDC2006T13)[5]. By ordering them according to the occurrence frequencies, we collected top 3,000 expressions. We then deleted all the MWEs already known as fixed

[1] http://typo.uni-konstanz.de/parseme/

[2] http://www.allinone-english.com/A13E/phrases-table-A-K.html, /phrases-table-L-Z.html

[3] http://ejje.weblio.jp/cat/dictionary/eidhg

[4] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

[5] https://catalog.ldc.upenn.edu/ldc2006t13

MWEs or phrasal verbs based on the lexicons constructed by previous works, (Shigeto et al., 2015) and (Komai et al., 2015). This results in 2,927 MWEs, which are our starting candidate flexible MWEs. At this stage, we do not know they are really flexible MWEs. Moreover, even if a candidate MWE is known as a flexible MWE, we do not know how flexible it is, that is, what kind of modifications it can involve.

4 Identification and Annotation of MWEs with Dependency Structure

4.1 Overview and objective

One of our objectives is to construct an English flexible MWE lexicon with the information of the degree of flexibility. Here, we only focus on flexibility concerning modification within the expression. For example, an MWE "a number of" can be used as "a growing number of" or "a very large number of". Finding those occurrences and their syntactic uniformity, we can guess that "a number of" can involve a word that modifies "number" in the expression[6]. To know correct syntactic structure of candidate MWEs, we make use of the Wall Street Journal portion of OntoNotes Release 5.0 (LDC2013T19) and converted all the phrase structure trees into dependency structure trees (those based on Stanford dependency[7]). The reason we used dependency tree rather than phrase structure trees is that the phrase structures in Peen Treebank are not uniform on their structure and phrase names. The same MWEs or the phrases that include them are in places annotated in slightly different phrase structures or with different phrase names. When they are converted into dependency structures, they become quite uniform.

Another objective is to annotate all the occurrences of MWEs in OntoNotes both in fixed or flexible forms. For all the possible occurrences of an MWE, that is, the occurrences of not only the exact appearances of the MWE but also the appearances that have one or more words intervened in the expression, we made annotation. With the help of dependency information obtained from the phrase structure tree, we semi-automatically annotated correct occurrences of MWEs. The same forms of some MWEs can be in literal usage. So, we are going to manually check all the annotation results before making them open to public.

The following subsections explain how we conducted the semi-automatic annotation of MWE candidates.

4.2 Extraction of Dependency Structure that Cover MWEs

This section describes the method for extracting dependency tree fragments that cover candidate MWEs. We used the Wall Street Journal portion (wsj_00-24) of OntoNotes after converting the phrase structure trees into Stanford dependency trees.

We took the following steps:

1. For each candidate MWE, we first extract all the sentences in OntoNotes that include the MWE in a flexible form. Fo example, in the case of "a number of", we extract all the sentences contain "a", "number" and "of" in this ordering. This process extracts quite a large number of sentences, but captures all sentences that potentially include the MWE.

2. We convert all the sentences into Stanford dependency (de Marneffe and Manning, 2008)[8].

3. For each sentence, we extract the minimal dependency subtree that covers the all the words comprising the MWE. An example of an extracted subtree is shown in Figure 1.

4. Within the subtree, there can be some other words or subtree that do not comprise the MWE. In the above case, the subtree consisting of "division heads" is an example. In such a case, we leave only the head of the subtree and delete all other children. Then we replace all the words that do not comprise the MWE with the POS labels. In the above example, we obtain the tree that represents a flexible usage of the MWE, "*a JJ number of NN*" (shown in Figure 2).

[6]The example "a very large number of" includes two words between "a" and "number", while only "large" modifies "number".

[7]http://nlp.stanford.edu/software/stanford-dependencies.shtml

[8]We designated "-conllx -basic -makeCopulaHead - keepPunct" as an option for the conversion command

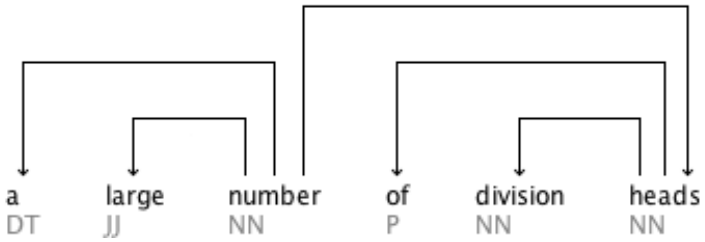

Figure 1: Minimal dependency subtree that covers “a number of”

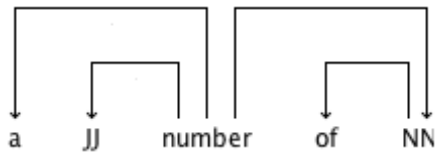

Figure 2: Representation of flexible usage of “a number of’

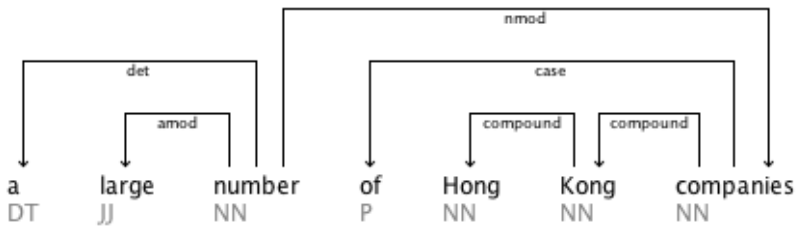

Figure 3: Another Minimal Dependency Subtree of “a number of”

Figures 3, 4, and 5 show three occurrences of “a number of” in different forms (all the figures show the minimal subtrees that include this MWE).

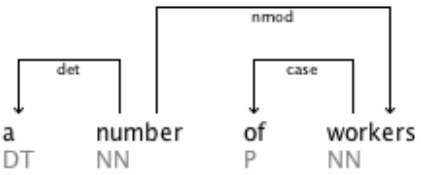

Figure 4: Subtree of ”a number of workers ...”

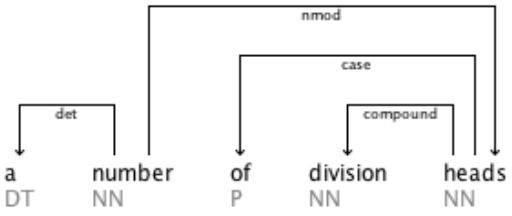

Figure 5: Subtree of ”a number of division heads ...”

After the procedure described above, we obtain isomorphic trees with only difference of existence of modifiers within the expression, e.g., the tree in Figure 3 includes a JJ as a modifier of “number”. From those trees we can obtain the dependency tree shown in Figure 6 as a flexible MWE so that “number” can have an internal modifier. In the figure, *1 is a wild card to be defined as a JJ or an empty element in the current case, but will be eventually defined as $\{\epsilon$, JJ, NN, VBG, VBN$\}$ and *2 is defined as $\{$NN, SYM$\}$ since words with those POS tags appear at the corresponding positions in some examples.

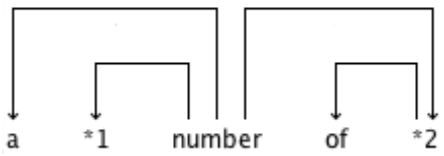

Figure 6: Representation of flexible MWE “a number of”

For each candidate MWE, we run the above procedure and obtain all possible subtrees. Some of the subtrees are from the fixed form of the MWE, i.e., in the original sentences there are no extra words intervening the expression. Still, they do not necessarily produce the same subtree. Within those subtrees, we pick up the smallest one, and assume it as the dependency structure of the MWE in the usage of its fixed form and call it as the canonical tree of the MWE. We assume all the other cases as non-MWE usages of the expression. We extract all the subtrees extracted from flexible occurrences and compare them with the canonical tree. If they are isomorphic except for the structure stemming from additional words that appear within the MWE, we regard them as the true flexible usage of the MWE. For all the subtrees that are not the same as the canonical tree nor isomorphic to the canonical tree are regarded as false cases, meaning they are not the true usage of the MWE.

For the 2927 MWE candidates we collected, we looked for all the fixed and flexible occurrences of them in the total of 37,015 sentences. Only 1871 MWEs have at least one occurrence in the corpus. We then obtained 26,358 minimal subtrees, and 14,146 unique minimal subtrees. We summarize them in Table 1.

Number of MWE types	1871
Number of Minimal Subtrees	26,358
Number of Unique Subtrees	14,146

Table 1: Dependency Subtrees of MWEs obtained from OntoNotes

Those figures suggest and we confirmed that most of the false occurrences of MWEs are unique.

5 Automated Annotation of MWEs

By comparing the subtrees for each MWE, we apply the above mentioned process for identifying positive usages of the MWE and for obtaining the dependency tree representation of the MWE.

For each MWE candidate, the instances that correspond to the canonical dependency trees and those that produce its isomorphic dependency trees are regarded as positive and true usages of the MWE. We cannot make any decision on the MWEs that appear only once in the corpus. In the following analysis, we excluded those MWEs.

When we decided that the canonical usages and their isomorphic usages are positive usages of MWEs, we found 1,194 positive fixed cases (i.e., canonical usages), 1,704 positive flexible cases (i.e., isomorphic to canonical form), and 11,248 negative cases. Table 2 summarizes them.

label	count
Positive Fixed MWEs	1194
Positive Flexible MWEs	1704
Negative cases	11,248

Table 2: The number of Fixed and Flexible MW and examples

5.1 Some Problematic Examples

In this section, we show some examples that are difficult to discriminate based on the structure uniformity with canonical usages. Figure 7 shows a positive usage (i.e., the canonical usage) of “a certain”. Figure 8 shows a negative occurrence of this MWE. The dependency tree in Figue 8 is isomorphic to that in Figure 7 except for the existence of an adverb “almost” within the expression as a parent of “certain”. Although our procedure cannot identify the latter case as a negative example, it is clear that the head NN’s in the trees are at different positions in the latter expression. The head NN in the canonical usage appears to the right of “certain”, while the head NN in the negative case appears to the left of “certain”. Taking the relative positons of head or modifiers into consideration solves this problem. We are going to investigate if this is true in all other cases.

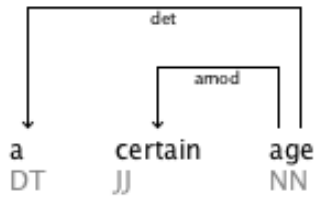

Figure 7: Canonical Subtree of ”a certain”

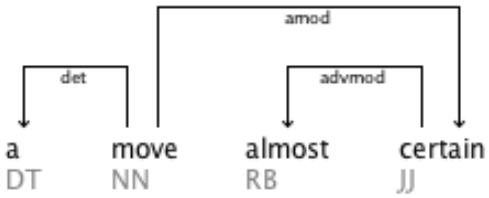

Figure 8: Negative Subtree of ”a certain”

Another problematic and difficult case is shown in Figures 9 and 10. Figure 9 shows the canonical usage of “a couple of”, and Figure 10 shows a variation of this MWE. Since the minimal subtree extracted from the latter example is not isomorphic to the former subtree, we cannot recognize this as a positive usage. On the other hand, if we like to regard the latter case as an admissible variation of the MWE “a couple of”, we need to find better ways for identifying these types of positive usages where an extra element is not necessarily a modifier (child) of a component of an MWE.

6 Conclusion and Feature work

We presented our ongoing project of English flexible multi-word expression lexicon construction and corpus annotation. We especially described a method of flexible MWE lexicon construction and their

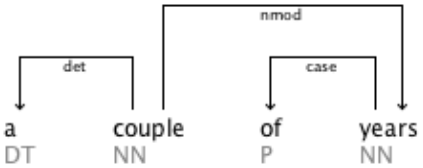

Figure 9: Subtree of "a couple of"

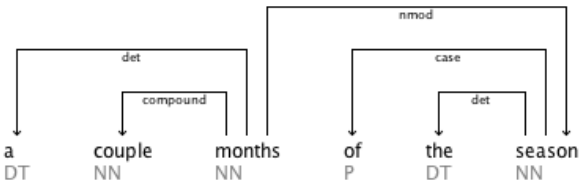

Figure 10: Subtree of "a couple month of"

annotation on a part of OntoNotes corpus. Our method enables semi-automatic annotation of flexible MWEs and also produces dependency structure representations of flexible MWEs.

While the method can achieve high recall of annotating positive occurrences in treebank, we need manual checking for those cases where the extracted minimal dependency subtrees are close but a slightly different from the canonical subtrees. Another problem we need to pursue is that the coverage of candidate MWEs is not wide enough. As we show in the experiments, within the MWE candidates we collected, only one third of them appear in the OntoNotes corpus. Furthermore, many of them show one or a small number of occurrences.

For the future work, we will try to collect far larger number of occurrences of the candidate MWEs in a large scale corpus, parse all the extracted sentences in dependency structure, and apply the method presented in this paper to those parsed results. Although the parsing accuracy is not 100%, handling a large number of examples hopefully provides results with high confidence.

Acknowledgement

This work was supported by CREST, JST, and JSPS KAKENHI Grant Number 15K16053.

References

Anne Abeillé, Lionel Clément, and Francois, Toussenel. 2003. *Building a Treebank for French*. In Treebanks : Building and Using Parsed Corpora. pages 165 – 188. Springer.

Marie Candito and Matthieu Constant. 2014. *Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing*. Proc. of ACL, pages 743 – 753.

Matthieu Constant and Joakim Nivre. 2016. *A Transition-Based System for Joint Lexical and Syntactic Analysis*. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 161 – 171.

de Marneffe, Marie-Catherine and Manning Christopher D. 2008. *The Stanford typed dependencies representation.*. In Proceedings of the Coling workshop on Cross-Framework and CrossDomain Parser Evaluation, pages 1 – 8.

Akihiko Kato, Hiroyuki Shindo, Yuji Matsumoto 2016. *Construction of an English Dependency Corpus incorporating Compound Function Words*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1667 – 1671.

Masayuki Komai, Hiroyuki Shindo, Yuji Matsumoto. 2015. *An Efficient Annotation for Phrasal Verbs using Dependency Information*. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC), Posters, pages 125 – 131.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. *OntoNotes: A unified relational semantic representation*. Proc. of ICSC, pages 517 – 526, Washington, DC, USA.

Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary and Petya Osenova 2016. *MWEs in Treebanks: From Survey to Guidelines*. Proc. of LREC-2016, pages 2323 – 2330, Portoro ž , Slovenia.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP*. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002), pages 1 − - 15, Mexico City, Mexico.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, Federico Sangati. 2015. *PARSEME – PARSing and Multiword Expressions within a European multilingual network*. Proceedings of the 7th Language & Technology Conference (LTC 2015), Pozna ń , Poland.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. *Comprehensive annotation of multiword expressions in a social web corpus*. Proceedings of LREC-2014, pages 455 – 461, Reykjavik, Iceland.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, Yuji Matsumoto. 2013. *Construction of English MWE Dictionary and its Application to POS Tagging*. Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 139 – 144.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer and Christopher D. Manning. 2014. *A Gold Standard Dependency Corpus for English*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).