Construction of English MWE Dictionary and its Application to POS Tagging

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, Yuji Matsumoto
Nara Institute Science of Technology (NAIST)
Ikoma, Nara 630-0192 Japan
yutaro-s@is.naist.jp

Abstract

This paper reports our ongoing project for constructing an English multiword expression (MWE) dictionary and NLP tools based on the developed dictionary. We extracted functional MWEs from the English part of Wiktionary, annotated the Penn Treebank (PTB) with MWE information, and conducted POS tagging experiments. We report how the MWE annotation is done on PTB and the results of POS and MWE tagging experiments.

1 Introduction

While there have been a great progress in POS tagging and parsing of natural language sentences thanks to the advancement of statistical and corpus-based methods, there still remains difficulty in sentence processing stemming from syntactic discrepancies. One of such discrepancies is caused by multiword expressions (MWEs), which are known and defined as expressions having "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002).

Sag et al. (2002) classifies MWEs largely into the following categories:

- Lexicalized phrases
  - fixed expressions: Those having fixed word order and form (e.g. *by and large*).
  - semi-fixed expressions: Those having fixed word order with lexical variation such as inflection, determiner selection, etc. (e.g. *come up with*).
  - syntactically flexible expressions: Those having a wide range of syntactic variability (e.g. phrasal verbs that take an NP argument between or following the verb and the particle).
- Institutionalized phrases
  - Phrases that are semantically and syntactically compositional, such as collocations (e.g. *traffic light*).

This paper reports our ongoing project for developing an English MWE dictionary of a broad coverage and MWE-aware natural language processing tools. The main contributions of this paper are as follows:

1. Construction of an English MWE dictionary (mainly consisting of functional expressions) through extraction from Wiktionary[1].
2. Annotation of MWEs in the Penn Treebank (PTB).
3. Implementation of an MWE-aware POS tagger and evaluation of its performance.

2 Related work

While there is a variety of MWE researches only a few of them focus on MWE lexicon construction. Though some examples, such as French adverb dictionaries (Laporte and Voyatzi, 2008; Laporte et al., 2008), a Dutch MWE dictionary (Grégoire, 2007) and a Japanese MWE dictionary (Shudo et al., 2011) have been constructed, there is no freely available English MWE dictionary with a broad coverage.

Moreover, MWE-annotated corpora are only available for a few languages, including French and

[1] https://en.wiktionary.org

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 139–144,
Atlanta, Georgia, 13-14 June 2013.

Swedish. While the British National Corpus is annotated with MWEs, its coverage is far from complete. Considering this situation, we started construction of an English MWE dictionary (with functional expressions first) and classified their occurrences in PTB into MWE or literal usage, obtaining MWE-annotated version of PTB.

The effect of MWE dictionaries have been reported for various NLP tasks. Nivre and Nilsson (2004) investigated the effect of recognizing MWEs in syntactic dependency parsing of Swedish. Korkontzelos and Manandhar (2010) showed performance improvement of base phrase chunking by annotating compound and proper nouns. Finlayson and Kulkarni (2011) reported the effect of recognizing MWEs on word sense disambiguation.

Most of the previous approaches to MWE recognition are based on frequency or collocation measures of words in large scale corpora. On the other hand, some previous approaches tried to recognize new MWEs using an MWE lexicon and MWE-annotated corpora. Constant and Sigogne (2011) presented MWE recognition using a Conditional Random Fields (CRFs)-based tagger with the BIO schema. Green et al. (2011) proposed an MWE recognition method using Tree Substitution Grammars. Constant et al. (2012) compared two phrase structure analysis methods, one that uses MWE recognition as preprocessing and the other that uses a reranking method.

Although MWEs show a variety of flexibilities in their appearance, most of the linguistic analyses consider the fixed type of MWEs. For example, the experiments by Nivre and Nilsson (2004) focus on fixed expressions that fall into the following categories:

1. Multiword names
2. Numerical expressions
3. Compound function words
   (a) Adverbs
   (b) Prepositions
   (c) Subordinating conjunctions
   (d) Determiners
   (e) Pronouns

Multiword names and numerical expressions behave as noun phrases and have limited syntactic functionalities. On the other hand, compound function words have a variety of functionalities that may affect language analyses such as POS tagging and parsing. In this work, we extract compound functional expressions from the English part of Wiktionary, and classify their occurrences in PTB into either literal or MWE usages. We then build a POS tagger that takes MWEs into account. In implementing this, we use CRFs that can handle a sequence of tokens as a single item (Kudo et al., 2004). We evaluate the performance of the tagger and compare it with the method that uses the BIO schema for identifying MWE usages (Constant and Sigogne, 2011).

3 MWEs Extraction from Wiktionary

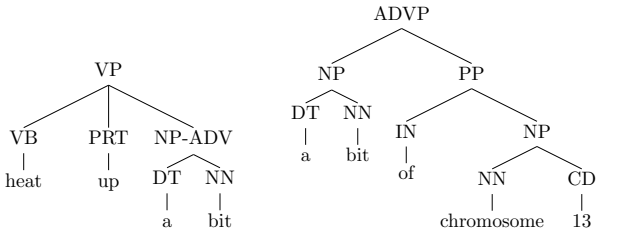
To construct an English MWE dictionary, we extract entries from the English part of Wiktionary (as of July 14, 2012) that include white spaces. We extract only fixed expressions that are categorized either as adverbs, conjunctions, determiners, prepositions, prepositional phrases or pronouns. We exclude compound nouns and phrasal verbs since the former are easily recognized by an existing method such as chunking and the latter need more sophisticated analyzing methods because of their syntactic flexibility. We also exclude multiword adjectives since many of them are semi-fixed and behave differently from lexical adjective, having predicative usage only. Table 1 summarizes the numbers of MWE entries in Wiktionary and the numbers of them that appear at least once in PTB.

4 Annotation of MWEs in PTB

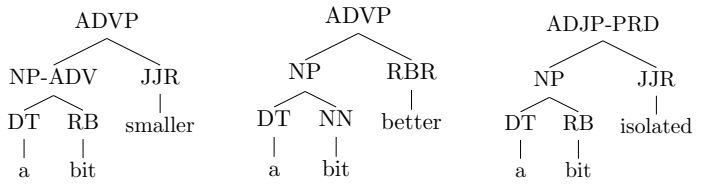
While it is usually not easy to identify the usage of an MWE as either an MWE or a literal usage, we initially thought that the phrase structure tree annotations in PTB would have enough information to identify their usages. This assumption is correct in many cases (Figures 1(a) and 1(b)). The MWE usage of "*a bit*" in Figure 1(a) is analyzed as "`NP-ADV`", suggesting it is used as an adverb, and the literal usage of "*a bit*" in Figure 1(b) is labeled as "`NP`", suggesting it is used literally. However, there are a number of examples that are annotated differently while their usages are the same. For example, Figures 1(c), 1(d) and 1(e) all show RB us-

Table 1: Number of MWE types in Wiktionary and Penn Treebank

	Adverb	Conjunction	Determiner	Preposition	Prepositional Phrase	Pronoun
Wiktionary	1501	49	15	110	165	83
PTB	468	35	9	77	66	18
Examples	*after all*	*as wll as*	*a number of*	*according to*	*against the law*	*no one*



(a) MWE usage as RB (b) Literal usage as NP



(c) MWE usage as RB (d) MWE usage as RB (e) MWE usage as RB

Figure 1: Examples of phrase structures annotated to "*a bit*"

age of "*a bit*" while they are annotated differently [2]. Sometimes, the same structure tree is annotated to instances of different usages (Figures 1(b) and 1(d)).

Therefore, for each MWE candidate, we first cluster its occurrences in PTB according to their phrase tree structures. Some of the clusters clearly indicate MWE usages (such as "`NP-ADV`" trees in Figures 1(a) and 1(c)). In such cases, we regarded all instances as MWE usages and annotated them as such. For inconsistent or ambiguous cases (such as "`NP`" trees in Figures 1(b), 1(d) and 1(e)), we manually classify each of them into either MWE or literal usage (some MWEs have multiple MWE usages). We find a number of inconsistent POS annotations on some internal words of MWEs (e.g. "*bit*" in Figures 1(c) and 1(e) are annotated as RB while they should be NN). We correct such inconsistent cases (correction is only done on internal words of MWEs, selecting the majority POS tags as correct). The total number of POS tag corrections made on PTB (chapter 00-24) was 1084.

[2]The POS tags in the trees are: RB(adverb), IN(preposition), DT(determiner), NN(common noun) ...

5 Experiments of POS tagging and MWE recognition

5.1 Experiment Setting

We conduct POS tagging experiments on the MWE-annotated PTB, using sections 0-18 for training and sections 22-24 for test as usual.

For the experiments, we use four versions of PTB with the following POS annotations.

(a) Original: PTB with the original POS annotation
(b) Revised: PTB with correction of inconsistent POS tags
(c) BIO MWE: MWEs are annotated with the BIO schema
(d) MWE: MWEs are annotated as single words

Concerning the MWE annotation in (c) and (d), the total number of MWE tokens in PTB is 12131 (9417 in the training chapters, 1396 in the test chapters, and 1319 for the remaining (development) chapters).

Each word is annotated with the following in-

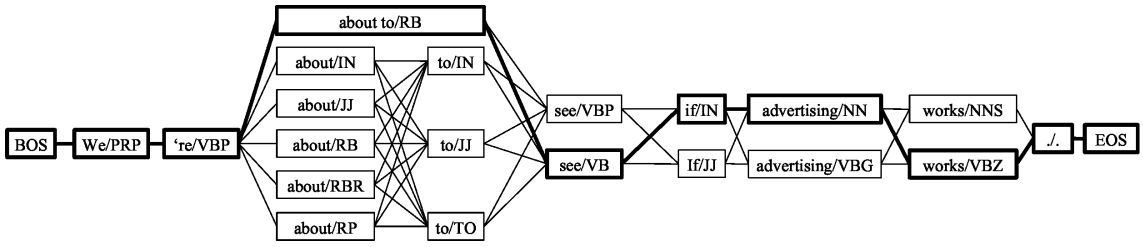


Figure 2: Example of lattice containing MWE ("*about to*/RB") (correct path is marked with bold boxes.)

Table 2: Examples of MWE annotations in four versions

Version	Word/POS
(a) Original	*about*/RB *to*/TO
(b) Revised	*about*/IN *to*/TO
(c) BIO MWE	*about*/RB-B *to*/RB-I
(d) MWE	*about to*/RB

formation: coarse-grained POS tag (CPOS), fine-grained POS tag (FPOS) and surface form. Each MWE is further annotated with its POS tag, surface form, its internal words with their POS tags.

Table 2 shows sample annotations of MWE "*about to*" in each of the four versions of PTB. In (a), "*about*/RB" is annotated incorrectly, which is corrected in (b). In (c), "-B" indicates the beginning token of an MWE and "-I" indicates an inside position of an MWE. In (d), "*about to*" is annotated as an RB (we omit the POS tags for its internal words, which are IN and TO).

We use a CRF-based tagger for training and test on all the four PTB versions. Our CRF can handle "words with spaces" (e.g. "*about to*" as a single token as well as separated tokens) as shown in Figure 2. This extension is only relevant to the case of the (d) MWE version.

Table 3 summarizes the set of feature templates used in the experiments. In Table 3, "Head POS" means the POS tag of the beginning token of an MWE. In the same way, "Tail POS" means the POS tag of the last token of an MWE. For example, for "*a lot of* /DT", its Head POS is DT and its Tail POS is IN.

We evaluate POS tagging accuracy and MWE recognition accuracy. In POS evaluation, each token receives a tag in the cases of (a), (b) and (c), so the tagging accuracy is straightforwardly calculated.

Table 3: Feature templates used in CRF training

Unigram features
Surface form
FPOS, Surface form
CPOS, Surface form
Bigram features (left context / right context)
Surface form / FPOS, Surface form
FPOS, Surface form / Surface form
Tail POS, Surface form / Head POS, Surface form
Surface form / Head POS
Tail POS / Head POS
Tail POS / Surface form

In the case of (d), since MWEs are analyzed as single words, they are expanded into the internal words with their POS tags and the evaluated on the token basis.

MWE recognition accuracy is evaluated for the cases of (c) and (d). For the purpose of comparison, we employ a simple baseline as well. This baseline assigns each occurrence of an MWE its most frequent usage in the training part of PTB. Evaluation of MWE recognition accuracy is shown in precision, recall and F-measure.

We use the standard set of features based on unigram/bi-gram of words/POS. For our MWE version, we add the word forms and POS tags of the first and the last internal words of MWEs as shown in Table 3.

5.2 Experimental Results

Table 4 shows the results of POS tagging. A slight improvement is observed in (b) compared with (a) because some of inconsistent tags are corrected. Further improvement is achieved in (d). The experiment on (c) does not show improvement even over

correct: · · · · who/WP after all/RB is/VBZ really/RB a/DT bit/JJ player/NN on/IN the/DT stage/NN · · · ·

system: · · · · who/WP *after/IN *all/DT is/VBZ really/RB *a bit/RB player/NN on/IN the/DT stage/NN · · · ·

Figure 3: Example of errors: "*after all* /RB" and "*a* /DT *bit* /JJ."

Table 4: Per token accuracy (precision)

Version	Accuracy
(a) Original	97.54
(b) Revised	97.56
(c) BIO MWE	97.32
(d) split MWE	97.62

Table 5: Recognition performance of MWEs

	Precision	Recall	F-measure
Baseline	78.79	80.26	79.51
(c) BIO	92.81	90.90	90.18
(d) MWE	95.75	97.16	96.45

(a). The reason may attribute to the data sparseness caused by the increased size of POS tags.

Table 5 shows the results of MWE recognition. Our MWE-aware CRF model (d) shows the best results. While the BIO model (c) significantly outperforms the baseline, it gives significantly lower results than our model.

We investigated errors in (d) and categorized them into three types.

- False Positive: System finds an MWE, while it is actually literal.
- False Negative: System misses to identify an MWE.
- Misrecognition: System finds an MWE wrongly (correct answer is another MWE).

Table 6 shows number of recognition errors of MWEs.

An example of the False Positive is "*a bit* /RB" in Figure 3, which actually is a literal usage and should be tagged as "*a* /DT, *bit* /NN".

An example of the False Negative is "*in black and white* /RB", which is not recognized as an MWE. One reason of this type of errors is low or zero frequency of such MWEs in training data. "*after all* /RB" (in Figure 3) is another False Negative example.

Table 6: Recognition error of MWEs

Error types	# of errors
False Positives	33
False Negatives	19
Misrecognition	17

One example of Misrecognition errors stems from ambiguous MWEs. For example, while "*how much*" only has MWE usages as RB, there are two RB usages of "*how much*" that have different POS tag sequences for the internal words. Other examples of Misrecognition are due to zero or low frequency MWEs, whose substrings also matches shorter MWEs: "*quite*/RB, *a few*/PRP" while correct analysis is "*quite a few*/RB", and "*the hell* /RB, *out of* /IN" while the correct analysis is "*the hell out of* /RB".

6 Conclusion and Future work

This paper presented our ongoing project for construction of an English MWE dictionary, and its application to MWE-aware POS tagging. The experimental results show that the MWE-aware tagger achieved better performance on POS tagging and MWE recognition. Although our current MWE dictionary only covers fixed types of functional MWEs, this dictionary and MWE annotation information on PTB will be made publicly available.

We plan to handle a wider range of MWEs such as phrasal verbs and other semi-fixed and syntactically flexible MWEs, and to develop a POS tagger and a syntactic parser on top of them.

References

Matthieu Constant and Anthony Sigogne. 2011. MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 49–56.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 204–212.

Mark Alan Finlayson and Nidhi Kulkarni. 2011. Detecting Multi-Word Expressions improves Word Sense Disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 20–24.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing *tour de force* with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 725–735.

Nicole Grégoire. 2007. Design and Implementation of a Lexicon of Dutch Multiword Expressions. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 17–24.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can Recognising Multiword Expressions Improve Shallow Parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 636–644.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '04, pages 230–237.

Eric Laporte and Stavroula Voyatzi. 2008. An Electronic Dictionary of French Multiword Adverbs. In *Language Resources and Evaluation Conference. Workshop Towards a Shared Task for Multiword Expressions*, MWE '08, pages 31–34.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A French Corpus Annotated for Multiword Nouns. In *Proceedings of the Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions*, MWE '08, pages 27–30.

Joakim Nivre and Jens Nilsson. 2004. Multiword Units in Syntactic Parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, MEMURA '04, pages 39–46.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann A Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. 2011. A Comprehensive Dictionary of Multiword Expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 161–170.