An Efficient Annotation for Phrasal Verbs using Dependency Information

Masayuki Komai **Hiroyuki Shindo** **Yuji Matsumoto**
Graduate School of Information and Science
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, 630-0192, Japan
{komai.masayuki.jy4, shindo, matsu}@is.naist.jp

Abstract

In this paper, we present an efficient semi-automatic method for annotating English phrasal verbs on the OntoNotes corpus. Our method first constructs a phrasal verb dictionary based on Wiktionary, then annotates each candidate example on the corpus as an either a phrasal verb usage or a literal one. For efficient annotation, we use the dependency structure of a sentence to filter out highly plausible positive and negative cases, resulting in a drastic reduction of annotation cost. We also show that a naive binary classification achieves better MWE identification performance than rule-based and sequence-labeling methods.

1 Introduction

Multiword Expressions (MWEs) are roughly defined as those that have "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2001). Vocabulary sizes of single words and MWEs have roughly the same size, thus MWE identification is a crucial issue for deep analysis of natural language text. Indeed, it has been shown in the literature that MWE identification helps various NLP applications, such as information retrieval, machine translation, and syntactic parsing (Newman et al., 2012; Ghoneim and Diab, 2013; Nivre and Nilsson, 2004). Since huge cost is necessary for annotation, there are few corpora that are sufficiently annotated for English MWEs. Schneider et al. (2014b) constructed an MWE-annotated corpus on English Web Treebank, and proposed a sequential labeling method for MWE identification. However, they tried to manually cover the types of comprehensive MWEs, and the number of instances for each MWE was very limited.

(a) We **bring** our computers **up**.
(b) She **goes over** the question.
(c) Someone goes over there.

Figure 1: a positive instance (a) of a separable expression "bring up", a positive instance (b) and a negative instance (c) of an inseparable expression "go over".

In this paper, we propose an efficient annotation method for separable MWEs appearing in a syntactic annotated corpus like the OntoNotes corpus. Although most of natural languages generally have a separable MWEs, an effort for separable MWE annotation is extremely limited. Therefore, we believe that constructing a large-scale corpus for separable MWEs is useful to develop and compare techniques of MWE identification. We especially focus on phrasal verbs that are a majority of separable MWEs, and propose an efficient method for phrasal verb annotation. To efficiently identify MWE usages, we exploit dependency structures on OntoNotes [1]. We also report experiments on MWE identification based on a binary classification, and show that it achieves better performance than rule-based and sequence-labeling methods. Further, we explore effective features for achieving high performance on the MWE identification task.

Our contributions are summarized as follows: (1)

[1] We use English OntoNotes corpus converted into the Stanford Dependency annotation format.

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 125 - 131
Shanghai, China, October 30 - November 1, 2015

Table 1: The number of MWE types.

	VB	RB	IN	JJ	PRP	DT	*Other*
types	994	395	94	17	14	12	6
example	go over	far from	in front of	ad hoc	anything else	a few	no way

We propose an efficient semi-automatic method for annotating phrasal verbs on OntoNotes. (2) We show that SVM-based naive classification is sufficient for accurate MWE identification. We also investigate effective features for MWE identification.

2 Related Work

MWEs can be roughly divided into two categories, separable and non-separable (or fixed) MWEs. Previous work annotated fixed MWEs on Penn Treebank, where they used syntactic trees of Penn Treebank and an MWE dictionary that is extracted from Wiktionary (Shigeto et al., 2013). In Schneider et al. (2014b), they annotated all types of MWEs on English Web Treebank completely by hand. Afterward, they added to supersenses, which mean coarse-grained semantic classes of lexical units (Schneider and Smith, 2015).

In MWE identification tasks, previous work integrated MWE recognition into POS tagging (Constant and Sigogne, 2011). An MWE identification method using Conditional Random Fields was also presented together with the data set (Shigeto et al., 2013). A joint model of MWE identification and constituency parsing was proposed (Constant et al., 2012). They allocated IOB[2] tags to MWEs and used MWEs as special features when reranking the parse tree. However, it is difficult for these methods to detect discontinuous MWEs.

In contrast, as for methods that can handle separable MWEs, Boukobza and Rappoport (2009) tackled MWE detection on specific MWE types with a binary classification method. In a framework of a sequential labeling method for MWE detection, a new IOB tag scheme, which is augmented to capture discontinuous MWEs and distinguish strong MWEs from week MWEs, was presented (Schneider et al., 2014a). Here strong MWEs indicate the expression which has strong idiomaticity, and week one indicate the expression which is to more likely to be a compositional phrase or collocation. Additionally, words between components of MWEs are called gaps, and the sequential labeling method that allocates IOB tags even to discontinuous sequences. This model is capable of capturing unknown MWEs, but it is difficult to detect new expressions with high accuracy.

3 Corpus Annotation

In this section, we present our annotation scheme for phrasal verbs. Our scheme mainly consists of three steps: acquisition of phrasal verbs, identification of phrasal verb occurrences on OntoNotes, and semi-automatic MWE classification with our heuristic rules.

3.1 Acquisition of Phrasal Verbs

First, we extract phrasal verb candidates from the English part of Wiktionary[3]. In particular, we parse a dump data of Wiktionary and extract verb entries that are composed of two or more words. We also collect phrasal verb candidates from the Web. For each MWE candidate, we manually check if they actually function as a phrasal verb. Moreover, we manually annotate whether their candidates are "separable" or "inseparable" and whether they are "transitive" or "intransitive". By "separable", we mean an object noun phrase can intervene between the main verb and a particle (e.g. **look** the tower **up**). Note that separable phrasal verbs do not always have an intervening object and also that inseparable phrasal verbs can be intervened by an adverb (e.g. **consist** largely **of**).

[2] **I**, **O** and **B** indicate **I**nside, **O**utside and **B**egin in a chunk respectively.

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

Table 2: Statistics in phrasal verb annotation

Annotation type	# instances
manual annotations	4022
automatic annotations	18574
Total	22596

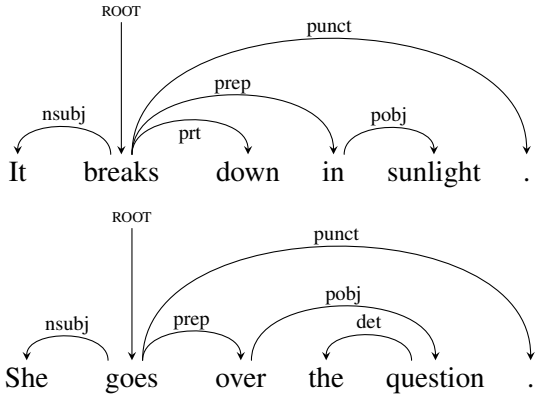

Figure 2: Examples of `prt` (break down) and `prep` (go over).

3.2 Identification of Phrasal Verb Occurrences in OntoNotes

Second, we retrieve all possible occurrences of phrasal verbs in OntoNotes. Here, we convert a surface word into a lemma form using Python-NLTK[4], then match the phrasal verb candidates with lemmatized words in OntoNotes. We regard this matching pattern as an instance. In Figure 1, instances (a) and (b) are positive instances where they are used as phrasal verbs. On the other hand, (c) is a negative instance where "go over" is not used as a phrasal verb but is used in the literal meaning. We extract discontinuous patterns as well since phrasal verbs have a potential of appearing discontinuously.

3.3 Semi-Automatic Annotation

Third, we check each instance whether it is used as a positive case or a negative one. Since it is too costly to check all instances manually, we propose to make use of dependency structures on OntoNotes to perform semi-automatic annotation. Indeed, we use the Stanford dependency converted from phrase structure trees on OnteNotes corpus. For each possible instance of a phrasal verb, we use the following relation between the verb and the particle that comprise the phrasal verb candidate: whether the verb and the particle appear adjacently or not and whether the verb and the particle have direct dependency or not, and if so the label of the dependency.

[4]http://www.nltk.org/

Table 3: Annotation rules for phrasal verb candidates. (In this table, "p" is positive, "n" is negative, "m" means manual annotation.)

direct dependency	continuous or not	dependency label	
True	$*$	`prt`	p
False	$*$	`prt`	n
True	True	`prep`	p
False	True	`prep`	m
True	False	`prep`	m
False	False	`prep`	n
True	$*$	*other*	m
False	$*$	*other*	n

Table 3 shows the whole annotation rules. In these rules, we especially focus on the dependency labels `prt` and `prep` in Stanford dependency (de Marneffe and Manning, 2008). The `prt` label, which directly connects a verb and a particle, may indicate the usage of a phrasal verb, and the `prep` label indicates a modifier to a verb as a prepositional phrase. Thus, we assume instances which have a direct `prt` dependency as positive instances. In the case of `prep`, there is a possibility of phrasal verbs or not. However, we assume instances which are adjacent, have a direct `prep` relation, and exist in our MWE lexicon as positive instances. If an instance is either not adjacent or having no direct relation with its particle or preposition, we put it for a candidate of manual checking. In this way, we have constructed annotation rules for MWE making full use of syntactic information.

For example, the instance of "break down" in Figure 2 has a direct relation with label `prt`, thus the first rule in Table 3 is applied. However, there are overlapping ambiguities that are not covered by these rules. For example, an instance "catch up with" can be labeled as positive, but another instance "catch up" of the part of "catch up with", may also be labeled as positive. When such an ambiguity oc-

Table 4: Corpus statistics of MWEs.

	# instances
positive instances	13214
negative instances	41167
Total	54381

Table 5: Evaluation of annotation rules.

Precision	Recall	F-value
62.72	82.60	71.30

curred, we manually checked their instances. As a result of annotation rules, we could reduce the cost of manual annotation considerably as shown in Table 2.

After annotating phrasal verbs on OntoNotes, we merge our annotation with the fixed MWE annotation done by (Shigeto et al., 2013). However, similar overlapping ambiguities have been generated again in this time (s.t. "get out of" and "out of"), we also manually eliminate these ambiguities.

In Table 4, we show statistics about our constructed corpus after merging. In total, 54381 instances are extracted from the 37015 sentences on OntoNotes,

4 Evaluation of Annotation Rule

In order to evaluate our annotation method, we also validate it on English Web Treebank annotated by (Schneider et al., 2014b). We first apply our method to English Web Treebank, then evaluate the quality of automatic annotation between automatically-annotated positive instances and gold MWEs on English Web Treebank. However, there is a large difference between both MWE candidates since annotators (the dictionary-based rule method and human) and domains of two corpora are different. In view of this, we evaluate only common phrasal verbs between two corpora.

In Table 5 we show the results of evaluating annotation rules. We obtain a sufficient recall, but the precision is lower than we expected. However, we consider this is unavoidable because annotators and domains are different as we have described preciously.

Table 6: The Feature list. W, G are the position list of the target MWEs to detect and of gaps. h and t are the position of the head MWEs and the tail. c_i, l_i and p_i is the ith context word, lemma and POS respectively. $[c_i]_j^k$ is the substring from jth to kth in c_i. $F(x)$ is the set that consisted of each element in the xth feature set.

basic features		
1	$\mathbf{c_i}, \mathbf{l_i}, \mathbf{p_i} \mid_{i \in W}$	
2	$c_i, l_i, p_i \mid_{i \in W}$	
3	$floor(\frac{\lvert G \rvert}{i})$	for i in $\{1, 2, 3, 4, 5\}$
context features		
4	$c_i, l_i, p_i \mid_{i=h-1}^{h-3}$	
5	$c_i, l_i, p_i \mid_{i=t+1}^{t+3}$	
6	$p_i \mid_{i \in G}$	
suffix & prefix features		
7	$[c_i]_1^j \mid_{j=1}^{3}$	for i in $h-1$ to $h-3$
8	$[c_i]_j^{\lvert c_i \rvert} \mid_{j=\lvert c_i \rvert-3}^{\lvert c_i \rvert}$	for i in $h-1$ to $h-3$
9	$[c_i]_1^j \mid_{j=1}^{3}$	for i in $t+1$ to $t+3$
10	$[c_i]_j^{\lvert c_i \rvert} \mid_{j=\lvert c_i \rvert-3}^{\lvert c_i \rvert}$	for i in $t+1$ to $t+3$
11	$[c_i]_1^j \mid_{j=1}^{3}$	for i in G
12	$[c_i]_j^{\lvert c_i \rvert} \mid_{j=\lvert c_i \rvert-3}^{\lvert c_i \rvert}$	for i in G
combination features		
13	$(e_1, e_2) \in \{F(1) \times F(2)\}$	
14	$(e_1, e_2) \in \{F(1) \times F(3)\}$	
15	$(e_1, e_2) \in \{F(1) \times F(4)\}$	
16	$(e_1, e_2) \in \{F(1) \times F(5)\}$	
17	$(e_1, e_2) \in \{F(1) \times F(6)\}$	
18	$(e_1, e_2) \in \{F(1) \times F(7)\}$	
19	$(e_1, e_2) \in \{F(1) \times F(8)\}$	
20	$(e_1, e_2) \in \{F(1) \times F(9)\}$	
21	$(e_1, e_2) \in \{F(1) \times F(10)\}$	
22	$(e_1, e_2) \in \{F(1) \times F(11)\}$	
23	$(e_1, e_2) \in \{F(1) \times F(12)\}$	

5 Experiments

In this section, we evaluate the performance of MWE identification task on our MWE-annotated OntoNotes. The MWE-annotated corpus used in our experiments contains fixed MWE annotations (Shigeto et al., 2013) and our phrasal verb annotations. The corpus is split into 2 sets: 33313 sentences (48970 instances) for training, and 3702 sentences (5411 instances) for testing. In these experiments, the system identifies MWEs given a sentence

Table 7: The experimental results.

	Precision	Recall	F-value
Rule-based method	62.93	**97.78**	76.58
Augmented IOB (Schneider et al., 2014a)	93.37	91.44	92.40
SVM	**93.77**	94.27	**94.02**

with gold POS.

5.1 Compared Methods

We compare SVM-based binary classification method against rule-based and sequential labeling method (Schneider et al., 2014a). The SVM method simply classifies each candidate instance as positive case or negative one. For the rule-based method, we use the following two simple rules. The first one is "if the target MWE is an instance of an inseparable phrasal verb and there is no gap between the verb and the particle (or preposition), then it is regarded as positive." The second one is "if the target MWE is an instance of a separable phrasal verb and the gap is 0 or equal to 1, then it is regarded as positive." Since our dictionary has information whether the target MWE is separable or not, we can use this information.

Table 6 shows the features that are used for SVM, which are categorized as four types: basic features, context features, suffix & prefix features, and combination features. In this table, bold $\mathbf{c}$, $\mathbf{l}$ and $\mathbf{p}$ are sequences that are concatenated context words, lemmas, POS sequences of target MWEs respectively. In respect to a classifier, we used SVM$_{light}$[5] with a linear kernel.

For sequential labeling method, we follow the previous work (Schneider et al., 2014a) for MWE identification. Their work exploits six types of tags, that is, {`O o B b I i`}, to handle with separable MWE identification, where **O**, **B**, **I** tags indicate **O**utside, **B**egin, **I**nside, and **o**, **b**, **i** tags indicate **o**utside, **b**egin, **i**nside in gaps respectively. In the experiments, we use their implementation[6] with exact match evaluation and set the recall-oriented hyperparameter ρ to 0.

[5]http://svmlight.joachims.org/

[6]http://www.cs.cmu.edu/~ark/LexSem/

Table 8: Investigation of effectiveness of features.

	F-value
basic features	92.89
+ **context features**	93.69
+ **suffix & prefix features**	93.06
+ **combination features**	94.02

5.2 Experimental Results

Table 7 summarizes the experimental results. In the table, we can see that SVM-based binary classification outperforms the rule-based and sequential labeling method. This result suggests that simple binary classification is sufficient for accurate MWE identification.

We also investigated which features are effective for our MWE identification task. Table 8 summarizes this analysis result. In the table, we can see that adding context features, suffix & prefix features, and those combinatorial ones to basic features successfully boost the identification performance. Further investigation of combinatorial features could be helpful for achieving better results, but we leave this for future work.

In error analysis, we found it is difficult for our method to detect the mutually-overlapping MWEs. For example, there should be the positive instance of "come out of" and the negative instance of "out of" in nature, but our model may say "positive" for both instances. Resolution of such conflicting cases should be investigated for future work.

Moreover, we have found that it is hard to recognize fixed MWEs, which appear continuously but are in literal usages. For example, "$a\ bit$" in "$is\ really\ a\ bit\ player\ on\ the\ stage$" is in literal usages. Our model tends to predict "positive" for such an instance.

6 Conclusion

We presented an semi-automatic method for annotating English phrasal verbs on the OntoNotes corpus. For efficient annotation, we use the dependency structure of a sentence to filter out positive and negative cases, resulting in a drastic reduction of annotation cost. We also reported that binary classification method outperformed rule-based and sequential labeling method. In order to improve the accuracy, we need a better model that takes wider contexts into consideration. We consider integration of syntactic parsing into MWE identification is one of such directions.

This paper also have described MWE annotation on OntoNotes. We will make the constructed dataset available on our website[7]. We are hoping that studies on MWEs are increased by using our dataset.

There are MWE types that we haven't handled at this work. For example, some flexible MWEs such as "take into account" are not annotated. Thus, we plan to annotate other discontinuous MWE types on OntoNotes so as to cover all MWEs on OntoNotes. We also believe that MWEs can include syntactic patterns, such as "not only ... but also". To deeply analyze a natural language text, we should explore such directions in future.

Acknowledgments

We thank the annotator Kayo Yamashita and the anonymous reviewers for their valuable comments. This work has been supported by JSPS KAKENHI Grant Numbers 15K16053 and 26240035. A part of this research was executed under the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Ram Boukobza and Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 468–477.

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)*, pages 204–212.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8.

Mahmoud Ghoneim and Mona Diab. 2013. Multiword expressions in the context of statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1181–1187.

David Newman, Nagendra Koilada, Jey Lau, and Tim Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *International Conference on Computational Linguistics (COLING)*, pages 2077–2092.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, pages 39–46.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, pages 1537–1547.

Nathan Schneider, Emily Danchik, Chris Dyer, and A. Noah Smith. 2014a. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association of Computational Linguistics (TACL) – Volume 2, Issue 1*, pages 193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013.

[7] http://cl.naist.jp/en/index.php?Code and Data

Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 139–144.