Construction of Large-scale English Verbal Multiword Expression Annotated Corpus

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

Abstract
Multiword expressions (MWEs) consist of groups of tokens, which should be treated as a single syntactic or semantic unit. In this work, we focus on verbal MWEs (VMWEs), whose accurate recognition is challenging because they could be discontinuous (e.g., take .. off). Since previous English VMWE annotations are relatively small-scale in terms of VMWE occurrences and types, we conduct large-scale annotations of VMWEs on the Wall Street Journal portion of English Ontonotes by a combination of automatic annotations and crowdsourcing. Concretely, we first construct a VMWE dictionary based on the English-language Wiktionary. After that, we collect possible VMWE occurrences in Ontonotes and filter candidates with the help of gold dependency trees, then we formalize VMWE annotations as a multiword sense disambiguation problem to exploit crowdsourcing. As a result, we annotate 7,833 VMWE instances belonging to various categories, such as phrasal verbs, light verb constructions, and semi-fixed VMWEs. We hope this large-scale VMWE-annotated resource helps to develop models for MWE recognition and dependency parsing that are aware of English MWEs. Our resource is publicly available.

Keywords: Multiword expressions, Phrasal verbs, Crowdsourcing

1. Introduction
Multiword expressions (MWEs) consist of groups of tokens, which should be treated as a single syntactic or semantic unit. MWEs are also known as “idiosyncratic interpretations that cross word boundaries” (Sag et al., 2002). In this paper, we focus on verbal MWEs (VMWEs) among various types of MWEs, such as compound nouns and compound function words. An accurate recognition of VMWEs is challenging because VMWEs could be discontinuous (e.g., take .. off). We show the main categories of VMWEs in Table 1.

While dependency parsing and MWE recognition could be solved independently, dependency structures in that each MWE is a syntactic unit are preferable to word-based dependency structures for downstream NLP tasks, such as semantic parsing. Because MWE recognition could help syntactic parsing (Nivre and Nilsson, 2004; Eryiğit et al., 2011), several works tackle MWE-aware dependency parsing in French (Candito and Constant, 2014; Nasr et al., 2015). They use French Treebank (Abeillé et al., 2003) because of its explicit MWE annotations.

Regarding English MWEs, Schneider et al. (2014) constructs an MWE-annotated corpus based on English Web Treebank (Bies et al., 2012). However, the number of VMWE occurrences (1,444) and types (1,155) in their corpus is relatively small-scale.

In this work, we conduct full-scale VMWE annotations on the Wall Street Journal (WSJ) portion of English Ontonotes (Pradhan et al., 2007), which results in 7,833 VMWE occurrences and 1,608 types. Concretely, we construct a VMWE dictionary based on the English-language Wiktionary 1. Based on this dictionary, we collect possible VMWE occurrences from Ontonotes and filter candidates with the help of gold dependency trees. To exploit crowdsourcing, we formalize VMWE annotations as a multiword sense disambiguation problem. This resource will enable the development of large-scale English MWE recognition and MWE-aware parsing models.

Our resource is publicly available at https://github.com/naist-cl-parsing/Verbal-MWE-annotations.

2. Corpus Construction
2.1. Candidate Extraction
First, we construct a VMWE dictionary by extracting multiword entries that have “English verbs” as categories.

Table 1: Main categories of VMWEs.

---

1https://en.wiktionary.org

2We select multiword entries that have “English_verbs” as categories.
Figure 1: Dependency trees with function-head (above) and content-head (below). We omit edges common in both trees. The box corresponds to a VMWE (“look at”). To filter a possible VMWE as a subtree of a dependency tree, a function-head scheme is preferable to a content-head scheme.

oneself). We exclude candidates that do not include any verbs by using gold part-of-speech information. Also, we filter out candidates that have other verbs or punctuation marks within the gaps.

Because most of the VMWEs are syntactically regular, we filter a VMWE whose components form a subtree in a Stanford basic dependency tree (Marneffe and Manning, 2008), which is converted from a phrase structure tree given in Ontonotes. We exploit Stanford basic dependency because its function-head scheme is suitable for filtering positive occurrences of VMWEs, that have a frequent POS pattern, “V IN”. In many cases, a noun phrase follows this type of MWEs. Therefore, in a content-head scheme like Universal Dependency (McDonald et al., 2013), a verb of this MWE governs a head of the noun phrase, that is, such MWE does not form a subtree (Figure 1). On the contrary, such MWE corresponds to a subtree in a function-head scheme.

Regarding phrasal verbs (PVs), we perform an additional filtering. In this work, we construct a VMWE-annotated corpus by extending Komai et al. (2015)’s corpus, because they have partially performed annotations of PVs in Ontonotes. For PVs that are not covered by their dictionary, we regard a candidate as a positive VMWE occurrence iff the dependency label is “prt”. For prepositional verbs, if the dependency label is “prep”, and there is no gap between the verb and the particle, we regard this candidate as a positive VMWE occurrence. This is subject to rules proposed by Komai et al. (2015). Otherwise, we conduct crowdsourced annotations.

2.2. Large-scale Annotations of VMWEs by Crowdsourcing

Based on the above filtering, we conduct large-scale VMWE annotations on the WSJ portion of English Ontonotes by crowdsourcing using a web interface shown in Figure 2. To exploit crowdsourcing, we formalize VMWE annotations as a multiword sense disambiguation problem. Annotators read a sentence in which components of a possible VMWE are highlighted. They are also given possible definitions of the VMWE, extracted from the English part of Wiktionary. For each VMWE, we provide one literal sense and non-literal senses. Based on this, they are asked to determine which definition most closely matches the meaning of highlighted words in the sentence. During annotations, workers are allowed to answer that the meaning of highlighted words is not in the given senses (“None of the above”), or they are not certain of the multiword sense (“Hard to judge”).

We collect crowdsourced annotations of VMWEs by using CrowdFlower. We set the following requirements: (1) Annotators belong to Level 3 contributors, who are regarded as the smallest group of most experienced, highest accuracy contributors on CrowdFlower. (2) Annotators live in countries with English as an official language. (3) Annotators achieve a success rate higher than 70 % in answering test questions, to which we give gold answers. To facilitate annotations, we provide workers with an interface to show multiple sentences (less than 6) that include possible occurrences of the same VMWE. Data collection costs $1,016 USD in total.

To determine whether each VMWE candidate is positive or not, we adopt the following criteria:

1. If all judgments correspond to the same sense, we

| # of constituent tokens | 2 | 3 | 4 | ≥ 5 | Total |
|-------------------------|---|---|---|-----|-------|
| VMWE instances          | 7,067 | 597 | 138 | 31 | 7,833 |

# of gaps | 0 | 1 | 2 |
|----------|---|---|---|
| VMWE instances | 6,855 | 968 | 10 |

VPCs, we regard a candidate as a positive VMWE occurrence iff the dependency label is “prt”. For prepositional verbs, if the dependency label is “prep”, and there is no gap between the verb and the particle, we regard this candidate as a positive VMWE occurrence. This is subject to rules proposed by Komai et al. (2015). Otherwise, we conduct crowdsourced annotations.

2.2. Large-scale Annotations of VMWEs by Crowdsourcing

Based on the above filtering, we conduct large-scale VMWE annotations on the WSJ portion of English Ontonotes by crowdsourcing using a web interface shown in Figure 2. To exploit crowdsourcing, we formalize VMWE annotations as a multiword sense disambiguation problem. Annotators read a sentence in which components of a possible VMWE are highlighted. They are also given possible definitions of the VMWE, extracted from the English part of Wiktionary. For each VMWE, we provide one literal sense and non-literal senses. Based on this, they are asked to determine which definition most closely matches the meaning of highlighted words in the sentence. During annotations, workers are allowed to answer that the meaning of highlighted words is not in the given senses (“None of the above”), or they are not certain of the multiword sense (“Hard to judge”).

We collect crowdsourced annotations of VMWEs by using CrowdFlower. We set the following requirements: (1) Annotators belong to Level 3 contributors, who are regarded as the smallest group of most experienced, highest accuracy contributors on CrowdFlower. (2) Annotators live in countries with English as an official language. (3) Annotators achieve a success rate higher than 70 % in answering test questions, to which we give gold answers. To facilitate annotations, we provide workers with an interface to show multiple sentences (less than 6) that include possible occurrences of the same VMWE. We collect three judgments for each of 2,135 possible VMWE occurrences. Data collection costs $1,016 USD in total.

To determine whether each VMWE candidate is positive or not, we adopt the following criteria:

1. If all judgments correspond to the same sense, we...
Table 4: VMWE statistics by POS patterns (for patterns occurring 10 or more times).

| POS pattern | Continuous | Discontinuous | Frequent MWEs |
|-------------|------------|---------------|---------------|
| V_IN        | 3,071      | 260           | base on : 142  
look for : 86  
focus on : 77  
go to : 70  
account for : 69 |
| V_RP        | 2,081      | 229           | set up : 62  
take over : 49  
point out : 47  
turn out : 43  
pick up : 39 |
| V_RB        | 547        | 116           | go back : 17  
come back : 17  
do well : 15  
down : 13  
look ahead : 13 |
| V_NN        | 280        | 167           | take place : 41  
do business : 27  
take effect : 26  
take steps : 24  
have time : 22 |
| V_DT_NN     | 114        | 45            | take a look : 13  
made a decision : 8  
pave the way : 5  
lay the groundwork : 5  
turn a profit : 5 |
| V_RP_IN     | 98         | 4             | come up with : 20  
make up for : 12  
keep up with : 8  
up to : 7  
add up to : 5 |
| V_JJ        | 77         | 11            | make sure : 14  
go wrong : 8  
go public : 6  
keep quiet : 5  
make much : 4 |
| V_IN_NN     | 56         | 26            | have in mind : 8  
take into account : 7  
set in motion : 5  
sign into law : 5  
take to heart : 4 |
| V_V         | 47         | 32            | be called : 34  
be had : 5  
have got : 4  
make known : 4  
let know : 4 |
| V PRP       | 77         | 0             | make it : 16  
have it : 10  
buy it : 6  
move it : 5  
find oneself : 5  
make much : 4 |
| V_PRPS_NN   | 49         | 1             | have one’s way : 5  
run one’s course : 4  
made one’s way : 3  
read someone’s lips : 3  
drag one’s feet : 2 |
| V_IN_IN     | 37         | 9             | get out of : 12  
come out of : 11  
made out of : 8  
grow out of : 4  
get through to : 1 |
| V_IN_DT_NN  | 33         | 11            | put on the block : 5  
come to an end : 5  
grind to a halt : 3  
jump on the bandwagon : 3  
get into the act : 3 |
| V_RB_IN     | 41         | 3             | get back to : 6  
send away from : 5  
cut back on : 4  
look back on : 4  
come up with : 3 |
| V_NN_IN     | 32         | 6             | take advantage of : 21  
take care of : 6  
keep tabs on : 3  
get wind of : 2  
take issue with : 1 |
| MD_V        | 17         | 6             | will do : 23  |
| V_DT_JJ_NN  | 17         | 0             | go a long way : 7  
look the other way : 4  
learn the hard way : 2  
take a back seat : 1  
fight a losing battle : 1 |
| V_DT_NN_IN  | 14         | 1             | keep a lid on : 3  
keep an eye on : 2  
put the brakes on : 2  
put the blame on : 1  
put a damper on : 1 |
| RB_V        | 5          | 7             | never mind : 4  
clear cut : 4  
second guess : 2  
reverse engineer : 1  
short circuit : 1 |
| V RP PRPS NN| 8          | 4             | make up one’s mind : 7  
pull in one’s horns : 2  
roll up one’s sleeves : 1  
clean up one’s act : 1  
hold up one’s end : 1 |

2.3 Resolution of Inclusions and Overlaps

Finally, we check inclusions and overlaps between annotations by us and those by (Komai et al., 2015), which results in 159 inclusions and 40 overlaps. Regarding inclusions, we adopt the broader MWE-spans. For instance, given two MWE occurrences corresponding to “come at” and “come at a price” in that a span of the latter includes a span of the former, we leave only the latter one. Concerning overlaps, we merge overlapped MWE-spans if we can get a new VMWE that is in both of the following dictionaries: Cambridge Dictionary 5 and The Free Dictionary 6. For instance, we get an occurrence of “take over the reins” by merging occurrences of “take the reins” and “take over”. Also, we resolve pseudo overlaps originating from false annotations. As a result, we reduce the number of overlaps to 11 instances, which correspond to essential overlaps, such as “look back” and “look ... on ... as” in the following sentence: “He may be able to look back on this election as the

5http://dictionary.cambridge.org  
6http://idioms.thefreedictionary.com
cies by the number of constituent word tokens (Table 7) prepositions, conjunctions, determiners, pronouns, or adverbs.

sifies their occurrences in Ontonotes into either MWE or MWEs. It constructs an MWE dictionary by extracting functional MWE annotations. It consists of phrase structure trees, augmented with morphological information and functional annotations of verbal dependents. Second, Baldwin et al. (2012) integrates annotations of these functional MWEs and named entities in Ontonotes by establishing MWEs as subtrees. They exploit this dataset for experiments on English MWE-aware dependency parsing.

4. Conclusion

In this work, we conduct large-scale annotations of English VMWEs in the Wall Street Journal portion of Ontonotes. Based on a VMWE dictionary extracted from English Wiktionary, we collect possible VMWE occurrences in Ontonotes, and filter candidates with the help of gold dependency trees. To take advantage of crowdsourcing, we formalize annotations of VMWEs as a multiword sense disambiguation problem. Our future work could involve the followings:

1. We plan to integrate our VMWE annotations into annotations for functional MWEs and named entities in Ontonotes by Kato et al. (2016) and Kato et al. (2017). This will help to develop models for MWE recognition and dependency parsing that are aware of various kinds of English MWEs.

2. We get VMWE occurrences in Ontonotes for only 1,608 out of 8,369 types in our VMWE dictionary. Therefore, we plan to explore VMWE occurrences on a larger corpus, such as the Annotated English Gigaword.

5. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 15K16053, 26240035 and JST CREST Grant Number JPMJCR1513, Japan. We are grateful to members of Computational Linguistics Laboratory at NAIST, and anonymous reviewers for valuable feedback. Regarding the preparation of a web interface for crowdsourcing, I am particularly grateful for the assistance given by Michael Wentao Li.

6. Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a Treebank for French, pages 165–187. Springer Netherlands, Dordrecht.

Baldwin, T., Kordoni, V., and Villavicencio, A. (2009). Prepositions in applications: A survey and introduction to the special issue. Computational Linguistics, 35(2):119–149.

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English web treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.

Candito, M. and Constant, M. (2014). Strategies for contiguous multiword expression analysis and dependency parsing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 743–753. Association for Computational Linguistics.

7. By functional MWEs, we mean MWEs that function either as prepositions, conjunctions, determiners, pronouns, or adverbs.

8. The NE annotations are given by Ontonotes.

9. http://catalog.ldc.upenn.edu/LDC2012T21
Eryiğit, G., Ilbay, T., and Can, O. A. (2011). Multiword expressions in statistical dependency parsing. In Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, pages 45–55, Dublin, Ireland, October. Association for Computational Linguistics.

Kato, A., Shindo, H., and Matsumoto, Y. (2016). Construction of an english dependency corpus incorporating compound function words. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may. European Language Resources Association (ELRA).

Kato, A., Shindo, H., and Matsumoto, Y. (2017). English multiword expression-aware dependency parsing including named entities. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 427–432, Vancouver, Canada, July. Association for Computational Linguistics.

Komai, M., Shindo, H., and Matsumoto, Y. (2015). An efficient annotation for phrasal verbs using dependency information. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 125–131.

Marneffe, M.-C. and Manning, D. C., (2008). Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, chapter The Stanford Typed Dependencies Representation, pages 1–8. Coling 2008 Organizing Committee.

McDonald, R., Nivre, J., Quirrbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 92–97. Association for Computational Linguistics.

Nasr, A., Ramisch, C., Deulofeu, J., and Valli, A. (2015). Joint dependency parsing and multiword expression tokenization. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1116–1126, Beijing, China, July. Association for Computational Linguistics.

Nivre, J. and Nilsson, J. (2004). Multiword units in syntactic parsing. In Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications, pages 39–46, Lisbon, Portugal.

Pradhan, S. S., Hovy, E. H., Marcus, M. P., Palmer, M., Ramshaw, L. A., and Weischedel, R. M. (2007). OntoNotes: A unified relational semantic representation. In Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA, pages 517–526.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (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, pages 1–15, London, UK, UK. Springer-Verlag.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The parseme shared task on automatic identification of verbal multiword expressions. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 31–47, Valencia, Spain, April. Association for Computational Linguistics.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 455–461, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1433.

Shigeto, Y., Azuma, A., Hisamoto, S., Kondo, S., Kouse, T., Sakaguchi, K., Yoshimoto, A., Yung, F., and Matsumoto, Y., (2013). Proceedings of the 9th Workshop on Multiword Expressions, chapter Construction of English MWE Dictionary and its Application to POS Tagging, pages 139–144. Association for Computational Linguistics.

Vincze, V. (2012). Light verb constructions in the szeged-paralelfix english-hungarian parallel corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). European Language Resources Association (ELRA).