Verb-Particle Constructions in Questions

Veronika Vincze[1,2]
[1]University of Szeged
Institute of Informatics
[2]MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

Abstract

In this paper, we investigate the behavior of verb-particle constructions in English questions. We present a small dataset that contains questions and verb-particle construction candidates. We demonstrate that there are significant differences in the distribution of WH-words, verbs and prepositions/particles in sentences that contain VPCs and sentences that contain only verb + prepositional phrase combinations both by statistical means and in machine learning experiments. Hence, VPCs and non-VPCs can be effectively separated from each other by using a rich feature set, containing several novel features.

1 Introduction

Multiword expressions (MWEs) contain more than one tokens but the whole unit exhibits syntactic, semantic or pragmatic idiosyncracies (Sag et al., 2002). Verb–particle constructions (VPCs), a specific type of MWEs, consist of a verb and a preposition/particle (like *set up* or *come in*). They often share their surface structures with compositional phrases, e.g. the phrases *to set up the rules* and *to run up the road* look similar but the first one contains a multiword expression while the other one is just a compositional phrase. This fact makes it hard to identify them on the basis of surface patterns. However, there are some syntactic or semantic processes that can be used to distinguish MWEs from compositional phrases. For instance, question formation (WH-movement), passivization and pronominalization are often listed among the distinctive tests (see e.g. (Kearns, 2002)). Phrasal-prepositional verbs usually employ the WH-words *what* or *who*, leaving the preposition at the end of the sentence as in *What did you set up*? In contrast, questions formed from compositional phrases usually contain the WH-words *where* or *when* as in *Where did you run?* However, the questions **Where did you set?* and **What did you run up?* are unacceptable.

In this study, we aim at investigating the behavior of verb-particle constructions in English questions. As a first step of our study, a database of questions will be created that contains verb-particle constructions and verb – prepositional phrase pairs. We will analyze these data from a quantitative point of view. This dataset will also constitute the training and test datasets for machine learning experiments. A rich feature set including morphological, semantic, syntactic and lexical features will be employed to learn the difference between verb-particle constructions and verb – prepositional phrase pairs in questions.

2 Related Work

Verb-particle constructions have been paid considerable attention in natural language processing. Baldwin and Villavicencio (2002) detected verb-particle constructions in raw texts on the basis of POS-tagging, chunking, statistical and lexical information. Kim and Baldwin (2006) relied on semantic information when detecting verb-particle constructions. Nagy T. and Vincze (2011) introduced a rule-based system using morphological features to detect VPCs in texts. Tu and Roth (2012) used syntactic and lexical features to classify VPCs candidates on a crowdsourced corpus. Nagy T. and Vincze (2014) implemented VPC-Tagger, a machine learning-based tool that selects VPC candidates on the basis of syntactic information and then classifies them as VPCs or not, based on lexical, syntactic and semantic features. Smith (2014) extracted VPCs from an English–Spanish parallel subtitles corpus.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 155–160,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

Here, we differ from earlier approaches in that we focus on just questions and we examine how linguistic features of questions may help in identifying VPCs in texts.

3 Data Collection

For data collection, we used three English corpora. First, we made use of the Google Web Treebank (Bies et al., 2012), which contains texts from the web annotated for syntactic (dependency) structures. Second, we used QuestionBank (Judge et al., 2006), which contains 4000 questions from two different sources: a test set for question-answering systems and a collection of question-answer type pairs. Each sentence in the treebank is assigned their constituency structures. Third, we used the Tu & Roth dataset (Tu and Roth, 2012), which contains verb-particle constructions and verb-prepositional phrase combinations.

From all three sources of data, we automatically filtered the sentences and selected questions from them. Furthermore, we also selected sentences that ended in a preposition or a particle (based on morphological information) and we grouped them into two classes: positive examples (questions with VPC) and negative examples (questions where the last token was a preposition due to preposition stranding). After these filtering steps, we got 280 questions out of which 227 were negative examples and the remaining 53 were positive examples. We parsed these sentences with the Bohnet dependency parser (Bohnet, 2010) in order to get a unified syntactic representation of the data. We will analyze these data from a quantitative point of view and report some statistics on them. This dataset will also be exploited by a machine learning system that aims at classifying each VPC candidate as a positive or negative one, which will be described in Section 5.

4 Statistical data

Here we will show some statistical data on the distribution of verbs, particles and WH-words in our dataset. We emphasize that our dataset is small and thus our results should be interpreted as showing only particular tendencies, and they should not be generalized.

4.1 Verbs

We first investigated what the distribution of the most frequent verbs are in the data. Table 1 shows the results, which are significant (χ^2-test, p = 6.72297E-12).

	positive	**negative**	**total**
be	0	36	36
come	5	18	23
get	10	2	12
go	3	2	5
grow	3	0	3
look	1	2	3
make	6	19	25
set	1	0	1
stand	0	32	32
take	2	2	4
turn	1	1	2
other	21	113	134

Table 1: Distribution of verbs.

The data reveal that there are some interesting differences in the distribution of verbs. For instance, it is a small set of verbs that can occur in positive examples (i.e. as part of a VPC), and there are verbs that occur exclusively as negative examples in the data such as *be* or *stand.*

4.2 Prepositions

We also analyzed the distribution of prepositions in positive and negative sentences. The results are shown in Table 2. Again, the results are significant (χ^2-test, p = 5.50637E-30). As can be seen, a small set of prepositions is responsible for most of the positive data. On the other hand, there are prepositions that do not occur in verb-particle constructions (at least in this dataset). Thus, the preposition itself seems to be a good indicator whether the construction is a genuine VPC or not.

Having a closer look at directional prepositions (marked with bold in Table 2), i.e. prepositions the meaning of which is related to spatial movement, a similar picture can be drawn. The prepositions *down*, *out* and *up* usually occur as parts of VPCs while *in* and *into* usually occur as parts of prepositional phrases. Results are significant (χ^2-test, p = 3.16905E-15).

The dependency labels of the prepositions are shown in Table 3. Results are significant here as well (χ^2-test, p = 9.58168E-09). Table 4 illustrates whether the preposition had any dependents in the syntactic tree and if yes, what its label was. Results are significant (χ^2-test, p = 0.0234).

	positive	**negative**	**total**
about	2	3	5
along	0	1	1
by	1	3	4
down	3	0	3
for	0	61	61
in	5	72	77
into	0	6	6
off	4	1	4
on	8	12	20
out	15	2	17
over	0	1	1
through	1	0	1
to	0	4	4
up	12	1	13
other	2	61	63

Table 2: Distribution of (**directional**) prepositions.

	positive	**negative**	**total**
ADV	12	110	122
PRT	37	59	96
other	4	58	62

Table 3: Dependency labels of prepositions. ADV: adverbial modifier, PRT: particle.

	positive	**negative**	**total**
COORD	1	0	1
PMOD	7	57	64
no child	45	170	215

Table 4: Dependency labels of the dependents of prepositions. COORD: coordination, PMOD: prepositional modifier.

	positive	**negative**	**total**
how	9	3	12
what	10	193	203
when	8	1	9
where	6	14	20
whom	1	0	1
why	2	0	2
which	0	3	3
who	0	5	5
other	17	8	25

Table 5: WH-words.

	positive	**negative**	**total**
WDT	4	8	12
WP	10	194	204
WRB	24	18	42
other	15	7	22

Table 6: POS codes of WH-words. WDT: WH-determiner, WP: WH-pronoun, WRB: WH-adverb.

4.3 WH-words

We also investigated the distribution of WH-words in the data. As can be seen from Tables 5 and 6, both WH-words and their morphological codes show significant differences between positive and negative sentences (χ^2-test, p = 2.89581E-25 for WH-words, p = 4.45435E-22 for codes). As for their dependency labels (see Table 7), question words functioning as adverbials of manner (MNR) and time (TMP) occur almost exclusively in sentences containing VPCs while when they function as subjects (SBJ), objects (OBJ) or arguments of prepositions (PMOD), the sentence usually does not contain a VPC. Results are significant (χ^2-test, p = 1.42263E-10).

5 Machine Learning experiments

We also carried out some machine learning experiments on the data. We implemented some of the features used by Nagy T. and Vincze (2014) and based on their results, we trained a J48 model (Quinlan, 1993) and an SVM model (Cortes and Vapnik, 1995) on the data (using Weka's (Hall et al., 2009) default settings) applying ten fold cross validation. As an evaluation metric, we used accuracy score. We use majority labeling as a baseline result, which yields an accuracy score of 81.07%.

	positive	negative	total
ADV	3	0	3
LOC	2	1	3
MNR	7	0	7
OBJ	2	9	11
PMOD	0	56	56
SBJ	8	51	59
TMP	8	1	9
other	15	56	71

Table 7: Dependency labels of WH-words. ADV: adverbial modifier, LOC: adverbial modifier of location, MNR: adverbial modifier of manner, OBJ: direct object, PMOD: prepositional modifier, SBJ: subject, TMP: adverbial modifier of time.

5.1 Feature Set

We made use of the following simple features:

WH-features: the WH-word; its POS code; whether it is sentence initial or not; its distance from the previous verb; its distance from the previous noun; its dependency label.

Verbal features: we investigated whether the lemma of the verb coincides with one of the most frequent English verbs since the most common verbs occur most typically in VPCs; we investigated whether the verb denotes motion as many verbs typical of VPCs express motion.

Prepositional features: whether the preposition coincides with one of the most frequent English prepositions; whether the preposition denotes direction; whether the preposition starts with *a* since etymologically, the prefix *a* denotes motion (like in *across*); its position within the sentence; its dependency label; whether the preposition has any children in the dependency tree.

Sentence-level features: the length of the sentence; we noted if the verb and the preposition both denoted motion or direction since these combinations usually have compositional meaning (as in *go out*); whether the verb had an object in the sentence; whether a pronominal object occurred in the sentence; whether a pronominal subject occurred in the sentence.

We note that WH-features and the last three of prepositional features are novel, which means that to the best of our knowledge, they have not been implemented in VPC detection yet.

5.2 Results

First, we trained our system with all the features, which resulted in an accuracy score of 90.36% with decision trees and 92.5% with SVM. Both results are well above our baseline (81.07%). Then we wanted to examine what the effect of the features that show significant differences can be on the results. Thus, we relied on the statistical results (see Section 4), and we retrained the system with only the statistically significant features, which are listed below:

1. the length of the sentence;
2. whether the verb and the preposition both denoted motion or direction;
3. the WH-word;
4. the POS code of the WH-word;
5. the dependency label of the WH-word;
6. whether the preposition coincides with one of the most frequent English prepositions;
7. whether the preposition denotes direction;
8. the position of the preposition within the sentence;
9. the dependency label of the preposition;
10. the dependency label of the preposition's child (if any);
11. whether the lemma of the verb coincides with one of the most frequent English verbs;
12. whether the verb and the preposition both denoted motion or direction.

With these settings, we could achieve an accuracy of 90% with decision trees and 92.14% with SVM, which is slightly worse than the previous results. Thus, the contribution of non-significant features is also important to the overall performance.

With further experiments, we found that the lexical features are the most important features for the system, as using only these features, accuracy scores of 89.64% and 93.93% can be obtained. Although our dataset is small, these results indicate that VPC detection can be relatively well performed with only a handful of features. All of our results are shown in Table 8.

	SVM	J48
baseline	81.07	81.07
all features	92.5	90.36
only significant features	92.14	90
only lexical features	93.93	89.64

Table 8: Results of machine learning experiments.

In order to test whether the same features can be applied to other datasets, we also experimented on the entire Tu & Roth dataset (i.e. we did not carry out any filtering steps). For the sake of comparability with previous results obtained for this corpus (Tu and Roth, 2012; Nagy T. and Vincze, 2014), here we applied an SVM model (Cortes and Vapnik, 1995) with 5 fold cross validation and obtained an accuracy score of 80.05%. On the same data, Tu and Roth (2012) obtained an accuracy score of 78.6%, which was outperformed by Nagy T. and Vincze (2014) with a score of 81.92%. Thus, our results can outperform those of Tu & Roth, but are below the one reported in Nagy T. and Vincze (2014). Thus, we can argue that our algorithm is capable of identifying VPCs effectively in a bigger dataset as well.

6 Discussion

Both our statistical investigations and machine learning experiences confirmed that the most important features in VPC detection are lexical features: i.e. the lemma of the verb, the preposition/particle and the WH-word can predict highly accurately whether the candidate is a VPC or not. Furthermore, semantic properties of the preposition – like denoting direction – also play a significant role. All these facts illustrate that relying on simple lexical features, VPC detection can be carried out effectively.

However, additional features that go behind a simple morphological analysis can also contribute to performance. For instance, investigating the dependency labels of the WH-word and the preposition reveals that there are significant differences among the positive and negative examples. It should be nevertheless noted that the dependency parser applies a separate label for VPCs, i.e. the particle is attached to the verb with the PRT relation, that is, the parser itself would also be able to identify VPCs (cf. Nagy T. and Vincze (2014)). However, as we can see from Table 3, the parser's performance is not perfect as it could achieve only an accuracy of 73.21% on our dataset and 58.13% on the Tu & Roth dataset. Thus, other features are also necessary to be included in the system.

Applying new features also contributed to the overall performance. We retrained our model on the Tu & Roth dataset without features that were implemented by us, in other words, we just applied features that had been introduced in earlier studies. In this way, we obtained an accuracy score of 77.46%, which means a gap of 3.81 percentage points. Thus, the added value of new features is also demonstrated.

7 Conclusions

In this paper, we investigated how verb-particle constructions behave in questions. We constructed a small dataset that contains questions and carried out statistical analyses of the data and also some machine learning experiments. From a statistical point of view, we found that there are significant differences in the distribution of WH-words, verbs and prepositions/particles in sentences that contain VPCs and sentences that contain only verb + prepositional phrase combinations. Dependency parsing also revealed some interesting facts, e.g. investigating whether the preposition has any children in the dependency tree proved also to be a significant factor. All these features proved useful in our machine learning settings, which demonstrated that VPCs and non-VPCs can be effectively separated from each other by using a rich feature set, containing several novel features. Our results achieved on a benchmark dataset are also very similar to those reported in the literature, thus the value of relying on additional features based on WH-words was also shown.

In the future, we would like to extend our database with additional examples and we plan to improve our machine learning system.

Acknowledgments

The work described in this paper has been supported by the IC1207 PARSEME COST action[1] and has been carried out during a Short Term Scientific Mission at Stanford University. The author is grateful to Dan Flickinger for his helpful comments and remarks on this paper.

[1] http://www.parseme.eu

References

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. Technical report, Linguistic Data Consortium, Philadelphia. LDC2012T13.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of Coling 2010*, pages 89–97.

Corinna Cortes and Vladimir Vapnik. 1995. *Support-vector networks*, volume 20. Kluwer Academic Publishers.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

John Judge, Aoife Cahill, and Josef Van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06)*, pages 497–504.

Kate Kearns. 2002. *Light verbs in English*. Manuscript.

Su Nam Kim and Timothy Baldwin. 2006. Automatic identification of English verb particle constructions using linguistic features. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pages 65–72.

István Nagy T. and Veronika Vincze. 2011. Identifying Verbal Collocations in Wikipedia Articles. In *Proceedings of the 14th International Conference on Text, Speech and Dialogue*, TSD'11, pages 179–186, Berlin, Heidelberg. Springer-Verlag.

István Nagy T. and Veronika Vincze. 2014. VPC-Tagger: Detecting Verb-Particle Constructions With Syntax-Based Methods. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 17–25, Gothenburg, Sweden, April. Association for Computational Linguistics.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

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 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15, Mexico City, Mexico.

Aaron Smith. 2014. Breaking bad: Extraction of verb-particle constructions from a parallel subtitles corpus. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 1–9, Gothenburg, Sweden, April. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth. 2012. Sorting out the Most Confusing English Phrasal Verbs. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 65–69, Stroudsburg, PA, USA. Association for Computational Linguistics.