Towards a Definition of Verb-Particle Constructions

Veronika Vincze
University of Szeged
vinczev@inf.u-szeged.hu

István Nagy T.
Collokia LLC
istvan@collokia.com

Introduction

Syntax-based detection of verb-particle constructions (VPCs) has been widely studied recently, especially in the English language. Here we briefly present the methods applied in earlier studies and we argue that differences in the definition of the concept of VPC and in annotation principles make it difficult to directly compare results obtained on different datasets and/or different methods. We also present a preliminary definition of VPC, which we hope to refine with the help of the PARSEME community.

VPC Detection

We briefly summarize syntax-based methods for VPC detection found in [4]. The special relation of the verb and particle within a VPC is distinctively marked in the Penn Treebank: the particle is assigned a specific part of speech tag (RP) and it also has a specific syntactic label (PRT). Thus, parsers trained on the Penn Treebank are expected to be able to identify VPCs in texts. Texts of the Wiki50 corpus [6] were parsed with the Stanford Parser [3] and the Bohnet parser [2] and if the parser correctly identified a PRT label, it was considered as a true positive. The Stanford parser achieved 91.09 (precision), 52.57 (recall) and 66.67 (F-measure) and the Bohnet Parser achieved 89.04 (precision), 58.16 (recall) and 70.36 (F-measure). Precision values are rather high but recall values are lower, which suggests that the sets of VPCs annotated in the Penn Treebank and Wiki50 may differ significantly.

A machine learning based approach is also presented in [4]. First, they syntactically parsed each sentence, and extracted potential VPCs with a syntax-based candidate extraction method. Afterwards, a binary classification was used to automatically classify potential VPCs as VPCs or not. For the automatic classification of candidate VPCs, they implemented decision trees with a rich feature set. This method achieved an F-score of 81.0, which outperformed results of the dependency parsers. The system was also evaluated on the Tu&Roth dataset [5], where it could obtain an accuracy of 81.92% and an F-score of 85.69.

Discussion

As we can see, there are differences in the performances of different methods on the same dataset on the one hand and differences in the performance of the same method on different datasets: machine learning methods outperform the parsers trained on the PENN Treebank and better results can be achieved on the Tu&Roth dataset than on the Wiki50 corpus.

We argue that the main reason behind differences is the lack of a unified definition of VPC and annotation principles. Although the PENN Treebank guidelines provide some hints on the annotation of VPCs, the low recall scores obtained by the parsers show that at least in the Wiki50 dataset, a wider set of linguistic constructions is annotated as VPCs than in the PENN treebank, which are unidentifiable by the parsers (i.e. they bear a dependency label different from PRT). Other studies such as [1] also emphasize that there are different definitions in use regarding VPCs, which, for instance, differ in the definition of particles (e.g. there is no consensus whether certain adjectives or verbs count as particles as in *cut short* or *let go*). Hence, differences in definitions may significantly affect annotation practice, leading to annotation discrepancies among corpora (and even among annotators of the same corpus).

Based on the above facts, we think that a revision of the VPC concept and definition is timely within the community. As a first step, we propose a preliminary and very general definition of VPCs that we intend to elaborate on with the PARSEME colleagues:

> A VPC is a verb + particle combination for which at least one of the following conditions holds:
>
> - its meaning is non-compositional (e.g. *do in*);
> - the verb and the particle can be separated with a noun or pronoun without any change in meaning (e.g. *set (it) up*);
> - there is an English synonym or a translational equivalent in another language which is a one-word unit or is a verb with a verbal prefix (e.g. *get away* and *escape*);
> - a noun can be derived from it (e.g. *breakthrough*).

A unified definition of VPCs can be exploited in corpus annotation and MWE detection as well as in lexicography and parsing. Besides, it also helps understand differences among annotated datasets and is also useful in error analysis of VPC detectors.

References

[1] Timothy Baldwin and Aline Villavicencio. 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, 2002. Association for Computational Linguistics.

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

[3] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Annual Meeting of the ACL*, volume 41, pages 423–430, 2003.

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

[5] Yuancheng Tu and Dan Roth. 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, 2012. Association for Computational Linguistics.

[6] Veronika Vincze, István Nagy T., and Gábor Berend. Multiword Expressions and Named Entities in the Wiki50 Corpus. In *Proceedings of RANLP 2011*, pages 289–295, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee.