Identifying verbal collocations in Wikipedia articles*

István Nagy T.[1] and Veronika Vincze[2]

[1] University of Szeged, Department of Informatics,
6720 Szeged, Árpád tér 2., Hungary
[2] MTA-SZTE Research Group on Artificial Intelligence,
6720 Szeged, Tisza Lajos krt. 103., Hungary
{nistvan,vinczev}@inf.u-szeged.hu

Abstract. In this paper, we focus on various methods for detecting verbal collocations, i.e. verb-particle constructions and light verb constructions in Wikipedia articles. Our results suggest that for verb-particle constructions, POS-tagging and restriction on the particle seem to yield the best result whereas the combination of POS-tagging, syntactic information and restrictions on the nominal and verbal component have the most beneficial effect on identifying light verb constructions. The identification of multiword semantic units can be successfully exploited in several applications in the fields of machine translation or information extraction.

Keywords: multiword expressions, verbal collocations, light verb constructions, verb-particle constructions, Wikipedia

1 Introduction

In natural language processing, the proper treatment of multiword expressions (MWEs) is essential for many higher-level applications (e.g. information extraction or machine translation). Multiword expressions are lexical items that can be decomposed into single words and display idiosyncratic features [11]. To put it differently, they are lexical items that contain space or 'idiosyncratic interpretations that cross word boundaries'. They are frequent in language use and because of their unique and idiosyncratic behavior, they often pose a problem to NLP systems.

In this work, we focus on various methods for detecting verbal collocations, i.e. verb-particle constructions (VPCs) and light verb constructions (LVCs) in Wikipedia articles. First, we offer a short description on characteristic features of these two types of multiword expressions, then related work is presented. Our methods are later described and results achieved are presented. The paper concludes with a discussion of results and future work.

* This work was supported in part by the National Innovation Office of the Hungarian government within the framework of the projects BELAMI and MASZEKER.

2 The characteristics of verb-particle constructions and light verb constructions

Verb-particle constructions contain a verb and a particle (usually a preposition), e.g. *kick off* or *set out*. They are also called phrasal or prepositional verbs and are highly characteristic of the English language, thus, they occur frequently in texts. The particle modifies the meaning of the verb: it may add aspectual information, may refer to motion or location or may totally change the meaning of the expression.

Light verb constructions are verb and noun combinations in which the verb has lost its meaning to some degree and the noun is used in one of its original senses (e.g. *have a walk* or *give advice*). They are usually distinguished from productive or literal `verb + noun` constructions on the one hand and idiomatic `verb + noun` expressions on the other hand in NLP literature: e.g. Fazly and Stevenson [7] use statistical measures in order to classify subtypes of verb + noun combinations and Diab and Bhutada [6] developed a chunking method for classifying multiword expressions.

Verbal collocations deserve special attention in NLP applications for several reasons. First, their meaning cannot be computed on the basis of the meanings of the parts of the collocation and the way they are related to each other (lack of total compositionality). Thus, the result of translating their parts literally can hardly be considered as the proper translation of the original expression. Second, light verb constructions (e.g. *make a mistake*) often share their syntactic pattern with literal verb + noun combinations (e.g. *make a cake*) or idioms (e.g. *make a meal*) and verb-particle constructions might follow the same pattern as a verb with a prepositional complement (*take on the task* or *sit on the chair*), which yields that their identification cannot be based on solely syntactic patterns. On the other hand, they are syntactically flexible, that is, they can manifest in various forms: the verb can be inflected, and the noun in light verb constructions can occur in its plural form or can be modified. The verbal component and the noun or the particle may not even be adjacent in e.g. passive sentences or with a pronominal object. However, for higher level applications (such as information extraction or machine translation) it is necessary to treat them as one unit, thus, their automatic identification is desirable.

3 Related work

In the following, methods developed for identifying light verb constructions and verb-particle constructions are summarized shortly.

Cruys and Villada Moirón [5] describe a semantic-based method for identifying verb-preposition-noun combinations in Dutch. Their method relies on selectional preferences for both the noun and the verb and they also make use of automatic noun clustering when considering the selection of semantic classes of nouns for each verb. Cook et al. [4] differentiate between literal and idiomatic usages of verb and noun constructions in English. Their basic hypothesis is that the canonical form of each construction occurs mostly in idioms since they show syntactic variation to a lesser degree than constructions in literal usage. Hence, they make use of syntactic fixedness of idioms when developing their unsupervised method. Bannard [3] seeks to identify verb and noun constructions in English on the basis of syntactic fixedness. He examines whether the noun can have a determiner or not, whether the noun can be modified and whether the construction can have a passive form, which features are exploited in the identification of the constructions. Samardžić and Merlo [12] analyze English and German light verb constructions in parallel corpora: they pay special attention to their manual and automatic alignment. They found that linguistic features (i.e. the degree of compositionality) and the frequency of the construction both have an effect on aligning the constructions.

Baldwin and Villavicencio [2] detect verb-particle constructions in raw texts. They make use of POS-tagging and chunking when developing their classifier while frequency and lexical information are also incorporated in their system. Kim and Baldwin [8] exploit semantic information when deciding whether verb-preposition pairs are verb-particle constructions or not. The (non-)compositionality of verb-particle combinations has been also paid attention in the literature. McCarthy et al. [10] implemented a method to indicate the compositionality of phrasal verbs and Baldwin [1] describes a dataset in which non-compositional VPCs can be found. Methods to extend the coverage of available VPC resources are proposed in [16].

4 Experiments

For the automatic identification of verbal collocations, we implemented several rule-based methods, which we describe below in detail.

4.1 Background

In order to identify multiword expressions, simple methods are worth examining, which can later serve as a basis for implementing more complex systems. Morphological information can be also exploited in the case of e.g. light verb constructions (the deverbal suffix of the noun may imply that it forms a light verb construction with the verb). Syntactic patterns can be also applied in identifying more complex or syntactically more flexible multiword expressions (e.g. some idioms can be passivized, compare *Who let the cat out of the bag?* and *The cat was let out of the bag*).

Although earlier studies on the detection of verbal collocations generally take syntactic information as a starting point (e.g. [4, 3, 10, 14]), that is, their goal is to classify constructions selected on the basis of syntactic patterns as literal or idiomatic, we would like to identify light verb constructions and verb-particle constructions in running text without assuming that syntactic information is necessarily available. Thus, in our investigations, we will pay distinctive attention to the added value of syntactic features on the system's performance. Given that we are not aware of any other corpora annotated for verb-particle combinations and light verb constructions at the same time, we restrict ourselves to rule-based methods since statistical methods would require a lot more data than available in our annotated database (see 4.3).

4.2 Methods for detecting verbal collocations

For identifying verbal collocations, we made use of several methods. In the case of 'POS-rules', each n-gram for which the pre-defined patterns (e.g. `VB.? (NN|NNS)` or `VB.? RP`) could be applied was accepted as LVC or VPC. For POS-tagging, we used the Stanford POS Tagger [15]. Since the methods to follow rely on morphological information (i.e. it is required to know which element is the verb, noun or particle), matching the POS-rules is a prerequisite to apply those methods.

The 'Suffix' method exploited the fact that many nominal components in light verb constructions are derived from verbs. Thus, in this case only constructions that matched our POS-rules and contained nouns ending in certain derivational suffixes were allowed.

The 'Most frequent' (MF) methods relied on the fact that the most common verbs occur typically in verbal collocations (e.g. *do*, *make*, *take*, *give* etc.) Thus, the 15 most frequent verbs (MFV) typical of light verb constructions and the 10 most frequent verbs typical of verb-particle combinations were collected and constructions that matched our POS-rules and where the stem of the verbal component was among those of the most frequent ones were accepted. The 20 most frequent particles (MFP) were similarly listed and the particle of the VPC candidate had to be among them.

The 'Stem' method pays attention to the stem of the noun. In the case of light verb constructions, the nominal component is typically one that is derived from a verbal stem (*make a decision*) or coincides with a verb (*have a walk*). In this case, we accepted only candidates that had a nominal component whose stem was of verbal nature, i.e. coincided with a stem of a verb.

Syntactic information can also be exploited in identifying MWEs. Typically, the syntactic relation between the verb and the nominal component in a light verb construction is `dobj` (using Stanford parser [9]) – if it is a prepositional light verb construction, the relation between the verb and the preposition is `prep`. The relation between a verb and its particle is `prt`. The 'Syntax' method accepts candidates among whose members the above syntactic relations hold.

We also combined the above methods to identify noun compounds and light verb constructions in our databases (the union of candidates yielded by the methods is denoted by $\cup$ while the intersection is denoted by $\cap$ in the respective tables). Rule-based methods were evaluated on our Wikipedia database and results are presented in 4.3.

4.3 Results

For the evaluation of our models, we developed a corpus of 50 Wikipedia articles, in which several types of multiword expressions (including verb-particle combinations and light verb constructions) and named entities were marked. The database contains 446 occurrences of verb-particle combinations and 368 occurrences of light verb constructions in 4350 sentences and can be downloaded under the Creative Commons license at `http://www.inf.u-szeged.hu/rgai/mwe`.

Results on the rule-based identification of light verb constructions can be seen in Table 1. The recall of the baseline (POS-rules) is high, however, its precision is low (i.e. not all of the candidates defined by the POS patterns are light verb constructions). The 'Most frequent verb' (MFV) feature proves to be the most useful: the verbal component of the light verb construction is lexically much more restricted than the noun, which is exploited by this feature. The other two features put some constraints on the nominal component, which is typically of verbal nature in light verb constructions: 'Suffix' simply requires the noun to end in a given n-gram (without exploiting further grammatical information) whereas 'Stem' allows nouns derived from verbs. When combining a verbal and a nominal feature, union results in high recall (the combinations typical verb + non-deverbal noun or atypical verb + deverbal noun are also found) while intersection yields high precision (typical verb + deverbal noun combinations are found only).

Method	**P**	**R**	**F**	**F with syntax**
POS-rules	7.02	76.63	**12.86**	16.56
Suffix	9.62	16.3	12.1	13.11
MFV	33.83	55.16	**41.94**	**45.31**
Stem	8.56	50.54	14.64	17.96
Suffix ∩ MFV	44.05	10.05	16.37	15.42
Suffix ∪ MFV	19.82	61.41	29.97	33.92
Suffix ∩ Stem	10.35	11.14	11.1	11.68
Suffix ∪ Stem	8.87	57.61	15.37	18.88
MFV ∩ Stem	39.53	36.96	38.2	39.81
MFV ∪ Stem	10.42	68.75	18.09	22.15
Suffix ∩ MFV ∩ Stem	**47.37**	7.34	12.7	11.96
Suffix ∪ MFV ∪ Stem	10.16	**72.28**	17.82	21.89

Table 1. Results of rule-based methods for light verb constructions in terms of precision (P), recall (R) and F-measure (F). POS-rules: matching of POS-patterns, Suffix: the noun ends in a given suffix, MFV: the verb is among the 15 most frequent light verbs, Stem: the noun is deverbal.

The added value of syntax was also investigated for LVC detection. As represented in the last column in Table 1, syntax clearly helps in identifying LVCs – except for two cases but its overall effect is to add up to 4% to the F-score. The best result, again, is yielded by the MFV method, which is about 30% above the baseline.

When identifying verb-particle constructions simply with POS-patterns, we get a baseline of 40.68 (F-score). Except for the 'Most frequent verb', each method can improve results as represented in Table 2. MFP proves to be the

Method	**P**	**R**	**F**
POS-rules	29.64	**64.8**	40.68
Syntax	91.89	53.36	67.52
POS ∩ syntax	92.12	49.78	64.63
MFV	41.96	10.54	16.85
MFV ∩ syntax	91.43	7.17	13.31
MFP	91.26	58.52	**71.31**
MFP ∩ syntax	93.59	49.1	64.41
MFV ∪ MFP	75.07	**60.09**	66.75
MFV ∪ MFP ∩ syntax	92.8	49.1	64.22
MFV ∩ MFP	**97.56**	8.97	16.43
MFV ∩ MFP ∩ syntax	96.97	7.17	13.36

Table 2. Results of rule-based methods for verb-particle constructions in terms of precision (P), recall (R) and F-measure (F). POS-rules: matching of POS-patterns, syntax: matching of syntactic patterns, MFV: the verb is among the 10 most frequent verbs, MFP: the particle is among the 20 most frequent particles.

best among the methods, which is due to the high precision of the method. When the two 'Most frequent' methods are contrasted, it is revealed that within a verb-particle construction, the particle seems to be lexically more restricted than the verb, thus, imposing constraints on the former leads to better results while the performance (especially recall) seriously declines when applying the MFV method. On the other hand, the intersection of the two methods yields the highest precision (constructions with a typical verb and a typical particle are only identified) while their union leads to the highest recall (except for the baseline method) since typical verb-atypical particle and atypical verb-typical particle pairs are also found besides typical verb-typical particle pairs.

The analysis of the added value of syntactic features reveals that although syntax proves to be the second best method, when combining it with other features, the overall performance of the system usually declines, however, precision improves a lot. This phenomenon might be connected to potential parsing errors where the parser fails to recognize the `prt` dependency relation between the particle and the verb, thus recall decreases. There is only one exception where the effect of syntax is the opposite: in the case of POS-rules, syntax obviously helps to identify VPCs. This can be expected since due to the common errors in POS-tagging, we chose to include particles and adverbs in our POS-patterns (the difficulties of distinguishing these types of parts of speech are also highlighted in [13]). Whereas this decision results in high recall values, precision seriously degrades, thus, a big pool of VPC candidates is yielded in this way from which the other methods (e.g. syntax) can select true positives.

5 Discussion

It is worth contrasting the results achieved for light verb constructions and verb-particle constructions. Making use of only POS-rules does not seem to be satisfactory for LVC detection. However, the most useful feature for identifying LVCs, namely, MFV proves to perform poorly for VPCs, which reflects that the verbal component of LVCs is lexically more restricted than the verbal part of VPCs. However, it is the particle in VPCs that is lexically more restricted as opposed to the verb, which is illustrated by the fact that the method MFP performs best.

As for light verb constructions, the feature 'Stem' seems to be beneficial for recall and this feature can be further enhanced since in some cases, the Porter stemmer did not render the same stem to derivational pairs such as *assumption* – *assume*, e.g. wordnet-based derivational information might contribute to performance.

Concerning syntactic information, it has clearly positive effects on LVC identification, however, its added value is not unequivocal in the case of verb-particle constructions. Due to possible parsing errors, syntactic features seem to introduce some noise in the performance of the system, thus, the combination of lexical and morphological features (POS-tagging) proves to be the most successful for identifying VPCs because of the relation between the two parts of the construction is rather lexical in nature (not syntactic in the sense that they constitute two separate phrases). On the other hand, light verb constructions do form a syntactic phrase (i.e. their parts can behave as separate phrases) hence syntactic features can be more successfully applied in their identification.

6 Conclusions

In this paper, we aimed at identifying verb-particle constructions and light verb constructions in running texts with rule-based methods and compared the effect of several features on performance. For verb-particle constructions, POS-tagging and restriction on the particle seem to yield the best result whereas the combination of POS-tagging, syntactic information and restrictions on the nominal and verbal component have the most beneficial effect on identifying light verb constructions. Special attention was paid to the role of syntax in identifying those types of multiword expressions: although it rather harms performance in the case of verb-particle constructions when combined with other features, it proves effective when applied alone and it is unambiguously helpful for identifying light verb constructions. As future work, we plan to further improve our methods by extending the set and scope of features and refining POS- and syntactic rules. We believe that detecting verb-particle constructions and light verb constructions (i.e. identifying multiword semantic units) can be successfully exploited in several applications in the fields of machine translation or information extraction.

References

1. Baldwin, T.: A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In: Proceedings of the LREC 2008 Workshop: Towards a

Shared Task for Multiword Expressions (MWE 2008), pp. 1–2. Marrakech, Morocco (2008)

2. Baldwin, T., Villavicencio, A.: Extracting the unextractable: a case study on verb-particles. In: Proceedings of the 6th Conference on Natural Language Learning, pp. 1–7. ACL (2002)
3. Bannard, C.: A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pp. 1–8. ACL, Prague (2007)
4. Cook, P., Fazly, A., Stevenson, S.: Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pp. 41–48. ACL, Prague (2007)
5. Van de Cruys, T., Villada Moirón, B.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pp. 25–32. ACL, Prague (2007)
6. Diab, M., Bhutada, P.: Verb Noun Construction MWE Token Classification. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pp. 17–22. ACL, Singapore (2009)
7. Fazly, A., Stevenson, S.: Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pp. 9–16. ACL, Prague (2007)
8. Kim, S.N., Baldwin, T.: Automatic identification of English verb particle constructions using linguistic features. In: Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, pp. 65–72 (2006)
9. Klein, D., Manning, C.D.: Accurate Unlexicalized Parsing. In: Proceedings of ACL 2003, pp. 423–430. ACL, Sapporo (2003)
10. McCarthy, D., Keller, B., Carroll, J.: Detecting a Continuum of Compositionality in Phrasal Verbs. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pp. 73–80. ACL, Sapporo (2003)
11. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) Proceedings of Conference on Intelligent Text Processing and Computational Linguistics 2002, pp. 1–15. Mexico City (2002)
12. Samardžić, T., Merlo, P.: Cross-Lingual Variation of Light Verb Constructions: Using Parallel Corpora and Automatic Alignment for Linguistic Research. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, pp. 52–60. ACL, Uppsala (2010)
13. Santorini, B.: Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, Department of Computer and Information Science, University of Pennsylvania (1990)
14. Tan, Y.F., Kan, M.-Y., Cui, H.: Extending corpus-based identification of light verb constructions using a supervised learning framework. In: Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts, pp. 49–56. ACL, Trento (2006)
15. Toutanova, K., Manning, C. D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of EMNLP 2000, pp. 63–70. ACL, Hong Kong (2000)
16. Villavicencio, A.: Verb-Particle Constructions and Lexical Resources. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pp. 57–64. ACL, Sapporo (2003)