Detecting noun compounds and light verb constructions: a contrastive study

Veronika Vincze[1], **István Nagy T.**[2] and **Gábor Berend**[2]
[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]Department of Informatics, University of Szeged
{nistvan,berendg}@inf.u-szeged.hu

Abstract

In this paper, we describe our methods to detect noun compounds and light verb constructions in running texts. For noun compounds, dictionary-based methods and POS-tagging seem to contribute most to the performance of the system whereas for light verb constructions, the combination of POS-tagging, syntactic information and restrictions on the nominal and verbal component yield the best result. However, focusing on deverbal nouns proves to be beneficial for both types of MWEs. The effect of syntax is negligible on noun compound detection whereas it is unambiguously helpful for identifying light verb constructions.

1 Introduction

Multiword expressions are lexical items that can be decomposed into single words and display idiosyncratic features (Sag et al., 2002; Calzolari et al., 2002; Kim, 2008). They are frequent in language use and they usually exhibit unique and idiosyncratic behavior, thus, they often pose a problem to NLP systems. A compound is a lexical unit that consists of two or more elements that exist on their own. 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*).

In this work, we aim at identifying *nominal compounds* and *light verb constructions* by using rule-based methods. Noun compounds belong to the most frequent MWE-classes (in the Wikipedia corpus we developed for evaluation (see 3.2), about 75% of the annotated multiword expressions were noun compounds) and they are productive, i.e. new nominal compounds are being formed in language use all the time, which yields that they cannot be listed exhaustively in a dictionary (as opposed to e.g. prepositional compounds). Their inner syntactic structure varies: they can contain nouns, adjectives and prepositions as well.

Light verb constructions are semi-productive, that is, new light verb constructions might enter the language following some patterns (e.g. *give a Skype call* on the basis of *give a call*). On the other hand, they are less frequent in language use (only 9.5% of multiword expressions were light verb constructions in the Wikipedia database) and they are syntactically flexible, that is, they can manifest in various forms: the verb can be inflected, the noun can occur in its plural form and the noun can be modified. The nominal and the verbal component may not even be adjacent in e.g. passive sentences.

Our goal being to compare how different approaches perform in the case of the different types of multiword expressions, we have chosen these two types of MWEs that are dissimilar in several aspects.

2 Related work

There are several applications developed for identifying MWEs, which can be classified according to the methods they make use of (Piao et al., 2003). First, statistical models rely on word frequencies, co-occurrence data and contextual information in deciding whether a bigram or trigram (or even an n-gram) of words can be labeled as a multiword expression or not. Such systems are used for several

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 116–121,
Portland, Oregon, USA, 23 June 2011.

languages and several types of multiword expressions, see e.g. Bouma (2010). The advantage of statistical systems is that they can be easily adapted to other languages and other types of multiword expressions, however, they are not able to identify rare multiword expressions (as Piao et al. (2003) emphasize, 68% of multiword expressions occur at most twice in their corpus).

Some hybrid systems make use of both statistical and linguistic information as well, that is, rules based on syntactic or semantic regularities are also incorporated into the system (Evert and Kermes, 2003; Bannard, 2007; Cook et al., 2007; Al-Haj and Wintner, 2010). This results in better coverage of multiword expressions. On the other hand, these methods are highly language-dependent because of the amount of linguistic rules encoded, thus, it requires much effort to adapt them to different languages or even to different types of multiword expressions. However, the combination of different methods may improve the performance of MWE-extracting systems (Pecina, 2010).

Several features are used in identifying multiword expressions, which are applicable to different types of multiword expressions to various degrees. Co-occurrence statistics and POS-tags seem to be useful for all types of multiword expressions, for instance the tool `mwetoolkit` (Ramisch et al., 2010a) makes use of such features, which is illustrated through the example of identifying English compound nouns (Ramisch et al., 2010b).

Caseli et al. (2010) developed an alignment-based method for extracting multiword expressions from parallel corpora. This method is also applied to the pediatrics domain (Caseli et al., 2009). Zarrieß and Kuhn (2009) argue that multiword expressions can be reliably detected in parallel corpora by using dependency-parsed, word-aligned sentences. Sinha (2009) detects Hindi complex predicates (i.e. a combination of a light verb and a noun, a verb or an adjective) in a Hindi–English parallel corpus by identifying a mismatch of the Hindi light verb meaning in the aligned English sentence. Van de Cruys and Moirón (2007) describe a semantic-based method for identifying verb-preposition-noun combinations in Dutch, which relies on selectional preferences for both the noun and the verb. Cook et al. (2007) differentiate between literal and idiomatic usages of verb and noun constructions in English. They make use of syntactic fixedness of idioms when developing their unsupervised method. Bannard (2007) also seeks to identify verb and noun constructions in English on the basis of syntactic fixedness. Samardžić and Merlo (2010) analyze English and German light verb constructions in parallel corpora. 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.

3 Experiments

In order to identify multiword expressions, simple methods are worth examining, which can serve as a basis for implementing more complex systems and can be used as features in machine learning settings. Our aim being to compare the effect of different methods on the identification of noun compounds and light verb constructions, we considered it important to develop methods for both MWE types that make use of their characteristics and to adapt those methods to the other type of MWE – in this way, the efficacy and the MWE-(in)dependence of the methods can be empirically evaluated, which can later have impact on developing statistical MWE-detectors.

Earlier studies on the detection of light verb constructions generally take syntactic information as a starting point (Cook et al., 2007; Bannard, 2007; Tan et al., 2006), that is, their goal is to classify verb + object constructions selected on the basis of syntactic pattern as literal or idiomatic. However, we do not aim at classifying LVC candidates filtered by syntactic patterns but at identifying them in running text without assuming that syntactic information is necessarily available. In our investigations, we will pay distinctive attention to the added value of syntactic features on the system's performance.

3.1 Methods for MWE identification

For identifying noun compounds, we made use of a list constructed from the English Wikipedia. Lowercase n-grams which occurred as links were collected from Wikipedia articles and the list was automatically filtered in order to delete non-English terms, named entities and non-nominal compounds etc. In the case of the method 'Match', a noun compound

candidate was marked if it occurred in the list. The second method we applied for noun compounds involved the merge of two possible noun compounds: if A B and B C both occurred in the list, A B C was also accepted as a noun compound ('Merge'). Since the methodology of dictionary building was not applicable for collecting light verb constructions (i.e. they do not function as links in Wikipedia), we could not apply these two methods to them.

In the case of 'POS-rules', a noun compound candidate was marked if it occurred in the list and its POS-tag sequence matched one of the previously defined patterns (e.g. `JJ (NN|NNS)`). For light verb constructions, the POS-rule method meant that each n-gram for which the pre-defined patterns (e.g. `VB.? (NN|NNS)`) could be applied was accepted as light verb constructions. For POS-tagging, we used the Stanford POS Tagger (Toutanova and Manning, 2000). Since the methods to follow rely on morphological information (i.e. it is required to know which element is a noun), matching the POS-rules is a prerequisite to apply those methods to identify MWEs.

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 contained nouns ending in certain derivational suffixes were allowed and for nominal compounds the last noun had to have this ending.

The 'Most frequent' (MF) method relied on the fact that the most common verbs function typically as light verbs (e.g. *do*, *make*, *take*, *have* etc.) Thus, the 15 most frequent verbs typical of light verb constructions were collected and constructions where the stem of the verbal component was among those of the most frequent ones were accepted. As for noun compounds, the 15 most frequent nouns in English were similarly collected[1] and the lemma of the last member of the possible compound 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 the nominal component / the last noun 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` or `prep` – using Stanford parser (Klein and Manning, 2003)). The relation between the members of a typical noun compound is `nn` or `amod` in attributive constructions. The 'Syntax' method accepts candidates among whose members these 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).

3.2 Results

For the evaluation of our models, we developed a corpus of 50 Wikipedia articles, in which several types of multiword expressions (including nominal compounds and light verb constructions) and Named Entities were marked. The database contains 2929 occurrences of nominal compounds and 368 occurrences of light verb constructions and can be downloaded under the Creative Commons licence at `http://www.inf.u-szeged.hu/rgai/mwe`.

Table 1 shows the results of our experiments. Methods were evaluated on the token level, i.e. each occurrence of a light verb construction had to be identified in text. It can be seen that the best result for noun compound identification can be obtained if the three dictionary-based methods are combined. We also evaluated the method of POS-rules without taking into account dictionary matches (POS-rules w/o dic), which result serves as the baseline for comparing the effect of LVC-specific methods on noun compound detection.

As can be seen, by adding any of the LVC-specific features, the performance of the system declines, i.e. none of them can beat the baseline. While the feature 'Stem' (and its combinations) improve precision, recall severely falls back: especially 'Most frequent noun' (MFN) has an extremely poor effect on it. This was expected since the lexical constraint on the last part of the compound heavily restricts the scope of the noun compounds available. On the

[1] as listed at `http://en.wikipedia.org/wiki/Most\_common\_words\_in\_English`

other hand, the 15 most frequent nouns in English are not derived from verbs hence they do not end in any of the pre-defined suffixes, thus, the intersection of the features 'MFN' and 'Suffix' does not yield any noun compound (the intersection of all the three methods also behaves similarly). It must be mentioned, however, that the union of all features yields the best recall as expected and the best F-measure can be achieved by the union of 'Suffix' and 'Stem'.

The effect of adding syntactic rules to the system is not unequivocal. In many cases, the improvement is marginal (it does not exceed 1% except for the POS-rules w/o dic method) or the performance even degrades. The latter is most obvious in the case of the combination of dictionary-based rules, which is mainly caused by the decline in recall, however, precision improves. The overall decline in F-score may thus be related to possible parsing errors.

In the case of light verb constructions, 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).

We also evaluated the performance of the 'Syntax' method without directly exploiting POS-rules. Results are shown in Table 2. It is revealed that the feature `dobj` is much more effective in identifying light verb constructions than the feature `prep`, on the other hand, `dobj` itself outperforms POS-rules. If we combine the `dobj` feature with the best LVC-specific feature (namely, MFV), we can achieve an F-measure of 26.46%. The feature `dobj` can achieve a recall of 59.51%, which suggests that about 40% of the nominal components in our database are not objects of the light verb. Thus, approaches that focus on only verb-object pairs (Cook et al., 2007; Bannard, 2007; Tan et al., 2006) fail to identify a considerable part of light verb constructions found in texts.

Method	**P**	**R**	**F**
Dobj	10.39	59.51	17.69
Prep	0.46	7.34	0.86
Dobj ∪ Prep	2.09	38.36	3.97
Dobj ∩ MFV	31.46	22.83	**26.46**
Prep ∩ MFV	3.24	5.12	4.06
Dobj ∪ Prep ∩ MFV	8.78	19.02	12.02

Table 2: Results of syntactic methods for light verb constructions in terms of precision (P), recall (R) and F-measure (F). Dobj: verb + object pairs, Prep: verb + prepositional complement pairs, MFV: the verb is among the 15 most frequent light verbs.

The added value of syntax was also investigated for LVC detection as well. As the results show, syntax clearly helps in identifying LVCs – 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.

4 Discussion

When contrasting results achieved for light verb constructions and noun compounds, it is revealed that the dictionary-based method applying POS-rules yields the best result for noun compounds and the MFV feature combined with syntactic information is the most useful for LVC identification. If no dictionary matches were taken into consideration, the combination of the features 'Suffix' and 'Stem' achieved the best result, however, 'Stem' alone can also perform similarly. Since 'Stem' identifies deverbal nouns, that is, nouns having an argument structure, it is not surprising that this feature is valuable in noun compound detection because the first part in the compound is most probably an argument of the deverbal noun (as in *noun compound detection* the object of *detection* is *noun compound*, in other words, we detect noun compounds). Thus, it will be worth examining how the integration of the 'Stem' feature can improve dictionary-based models.

Making use of only POS-rules does not seem to

Method	Noun compounds			NC + syntax			LVC			LVC + syntax		
	P	R	F	P	R	F	P	R	F	P	R	F
Match	37.7	54.73	44.65	49.64	48.31	48.97	-	-	-	-	-	-
Merge	40.06	57.63	47.26	51.69	47.86	49.70	-	-	-	-	-	-
POS-rules	55.56	49.98	52.62	59.18	46.39	**52.02**	-	-	-	-	-	-
Combined	59.46	52.48	**55.75**	62.07	45.81	**52.72**	-	-	-	-	-	-
POS-rules w/o dic	28.33	66.23	**39.69**	29.97	64.18	40.87	9.35	72.55	**12.86**	7.02	76.63	16.56
Suffix	27.02	8.91	13.4	28.58	8.84	13.5	9.62	16.3	12.1	11.52	15.22	13.11
MF	12.26	1.33	2.4	12.41	1.29	2.34	33.83	55.16	**41.94**	40.21	51.9	**45.31**
Stem	29.87	37.62	**33.3**	31.69	36.63	33.99	8.56	50.54	14.64	11.07	47.55	17.96
Suffix∩MF	0	0	0	-	-	-	44.05	10.05	16.37	11.42	54.35	18.88
Suffix∪MF	23.36	10.24	14.24	24.50	10.13	14.34	19.82	61.41	29.97	23.99	57.88	33.92
Suffix∩Stem	28.4	6.49	10.56	30.03	6.42	10.58	10.35	11.14	11.1	12.28	11.14	11.68
Suffix∪Stem	29.35	**40.05**	**33.87**	31.12	39.06	**34.64**	8.87	57.61	15.37	11.46	54.35	18.93
MF∩Stem	9.16	0.41	0.78	9.6	0.41	0.79	39.53	36.96	38.2	46.55	34.78	39.81
MF∪Stem	29.13	38.55	**33.18**	31.85	36.04	33.81	10.42	68.75	18.09	13.36	64.67	22.15
Suffix∩MF∩Stem	0	0	0	-	-	-	**47.37**	7.34	12.7	50.0	6.79	11.96
Suffix∪MF∪Stem	28.68	**40.97**	33.74	30.33	39.95	34.48	10.16	**72.28**	17.82	13.04	68.2	21.89

Table 1: Experimental results in terms of precision (P), recall (R) and F-measure (F). Match: dictionary match, Merge: merge of two overlapping noun compounds, POS-rules: matching of POS-patterns, Combined: the union of Match, Merge and POS-rules, POS-rules w/o dic: matching POS-patterns without dictionary lookup, Suffix: the (head) noun ends in a given suffix, MF: the head noun/verb is among the 15 most frequent ones, Stem: the (head) noun is deverbal.

be satisfactory for LVC detection. However, the most useful feature for identifying LVCs, namely, MFV/MFN proves to perform poorly for noun compounds, which can be explained by the fact that the verbal component of LVCs usually comes from a well-defined set of frequent verbs, thus, it is lexically more restricted than the parts of noun compounds. The feature 'Stem' helps improve 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*. For instance, derivational information encoded in wordnet relations might contribute to performance.

Concerning syntactic information, it has clearly positive effects on LVC identification, however, this influence is ambiguous in the case of noun compounds. Since light verb constructions form a syntactic phrase and noun compounds behave syntactically as one unit (having an internal syntactic hierarchy though), this result suggests that for noun compound detection, POS-tagging provides enough information while for light verb constructions, syntactic information is expected to improve the system.

5 Conclusions

In this paper, we aimed at identifying noun compounds and light verb constructions in running texts with rule-based methods and compared the effect of several features on detecting those two types of multiword expressions. For noun compounds, dictionary-based methods and POS-tagging seem to contribute most to the performance of the system whereas for light verb constructions, the combination of POS-tagging, syntactic information and restrictions on the nominal and verbal component yield the best result. Although the effect of syntax is negligible on noun compound detection, it is unambiguously helpful for identifying light verb constructions. Our methods can be improved by extending the set and scope of features and refining POS- and syntactic rules and they can be also adapted to other languages by creating language-specific POS-rules, lists of suffixes and light verb candidates.

For higher-level of applications, it is necessary to know which tokens form one (syntactic or semantic) unit, thus, we believe that our results in detecting noun compounds and light verb constructions can be fruitfully applied in e.g. information extraction or machine translation as well.

Acknowledgments

This work was supported in part by the National Innovation Office of the Hungarian government within the framework of the project MASZEKER.

References

Hassan Al-Haj and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of Coling 2010*, Beijing, China, August.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 1–8, Morristown, NJ, USA. ACL.

Gerlof Bouma. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 109–114, Uppsala, Sweden, July. ACL.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC-2002*, pages 1934–1940, Las Palmas.

Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. 2009. Statistically-driven alignment-based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 1–8, Singapore, August. ACL.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. 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*, pages 41–48, Morristown, NJ, USA. ACL.

Stefan Evert and Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Proceedings of EACL 2003*, pages 83–86.

Su Nam Kim. 2008. *Statistical Modeling of Multiword Expressions*. Ph.D. thesis, University of Melbourne, Melbourne.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, pages 49–56, Morristown, NJ, USA. ACL.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010a. Multiword Expressions in the wild? The mwetoolkit comes in handy. In *Coling 2010: Demonstrations*, Beijing, China, August.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010b. Web-based and combined language models: a case study on noun compound identification. In *Coling 2010: Posters*, Beijing, China, August.

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 CICLing-2002*, pages 1–15, Mexico City, Mexico.

Tanja Samardžić and Paola Merlo. 2010. 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*, pages 52–60, Uppsala, Sweden, July. ACL.

R. Mahesh K. Sinha. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 40–46, Singapore, August. ACL.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. 2006. 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*, pages 49–56, Trento, Italy, April. ACL.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP 2000*, pages 63–70, Stroudsburg, PA, USA. ACL.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 25–32, Morristown, NJ, USA. ACL.

Sina Zarrieß and Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30, Singapore, August. ACL.