Domain-Dependent Identification of Multiword Expressions

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

Abstract

The identification of different kinds of multiword expressions require different solutions, on the other hand, there might be domain-related differences in their frequency and typology. In this paper, we show how our methods developed for identifying noun compounds and light verb constructions can be adapted to different domains and different types of texts. Our results indicate that with little effort, existing solutions for detecting multiword expressions can be successfully applied to other domains as well.

1 Introduction

Multiword expressions (MWEs) are lexical units that consist of more than one orthographical word, i.e. a lexical unit that contains spaces (Sag et al., 2002; Calzolari et al., 2002). There are several methods developed for identifying several types of MWEs, however, different kinds of multiword expressions require different solutions. Furthermore, there might be domain-related differences in the frequency of a specific MWE type. In this paper, we show how our methods developed for identifying noun compounds and light verb constructions can be adapted to different domains and different types of texts, namely, Wikipedia articles and texts from various topics. Our results suggest that with simple modifications, competitive results can be achieved on the target domains.

2 Related work

There are several solutions developed for identifying different types of MWEs in different domains. Bonin et al. (2010) use contrastive filtering in order to identify multiword terminology in scientific, Wikipedia and legal texts: term candidates are ranked according to their belonging to the general language or the sublanguage of the domain. The tool `mwetoolkit` (Ramisch et al., 2010a) is designed to identify several types of MWEs in different domains, which is illustrated by identifying English compound nouns in the Genia and Europarl corpora and in general texts (Ramisch et al., 2010b; Ramisch et al., 2010c).

Statistical models are used for the identification of several types of multiword expressions in several languages (e.g. Bouma (2010),Villavicencio et al. (2007)). However, they require (costly) annotated resources on the one hand and they are not able to identify rare MWEs in corpora on the other hand – as Piao et al. (2003) emphasize, about 68% of multiword expressions occur only once or 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 (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 specific 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. Thus, the adaptation of linguistics-based models or hybrid models is required for identifying rare MWEs in small corpora from different domains.

3 Experiments

In this paper, we focus on the identification of two types of multiword expressions, namely noun compounds and light verb constructions. 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

Proceedings of Recent Advances in Natural Language Processing, pages 622–627,
Hissar, Bulgaria, 12-14 September 2011.

(e.g. *have a walk*).

We selected noun compounds since they are very frequent in language use (in the Wiki50 corpus (Vincze et al., 2011b) 67.3% of the sentences contain a noun compound on average). On the other hand, they are productive: new noun compounds are being created all the time hence they cannot be exhaustively listed. Light verb constructions are less frequent (8.5% of the sentences contain one), however, they are syntactically flexible: the nominal component and the verb may not be adjacent, which hinders their identification. Their proper treatment is especially important in information (event) extraction, where verbal elements play a central role and extracted events may differ if the verbal and the nominal component are not considered as one complex predicate.

For the automatic identification of noun compounds and light verb constructions, we implemented several rule-based methods, which we describe below in detail.

As opposed to earlier studies (Cook et al., 2007; Bannard, 2007; Tan et al., 2006), we would like to identify light verb constructions in running text without assuming that syntactic information is necessarily available (in line with Vincze et al. (2011a)). Thus, 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.

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 for identifying 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 matched our POS-rules and contained nouns that end in certain derivational suffixes were allowed.

The 'Most frequent verb' (MFV) method relied on the fact that the most common verbs function typically as light verbs (e.g. *do*, *make*, *take* etc.) Thus, the 12 most frequent verbs typical of light verb constructions 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 '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` or `partmod` (using Stanford parser (Klein and Manning, 2003)) – if it is a prepositional light verb construction, the relation between the verb and the preposition is `prep`. 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).

3.2 Corpora used for evaluation

For the evaluation of our models, we made use of three corpora. Data on the corpora are shown in Table 1.

First, we used Wiki50 (Vincze et al., 2011b), in which several types of multiword expressions (including nominal compounds and light verb constructions) and named entities were marked. The corpus contains 2929 occurrences of nominal compounds and 368 occurrences of light verb con-

Corpus	Sentence	Token	NC	LVC
Wikipedia	4350	114,570	2929	368
BNC dataset	1000	21,631	368	-
Parallel	14,262	298,948	-	1100

Table 1: Corpora used for evaluation. NC: noun compounds, LVC: light verb constructions.

structions.

Our methods for identifying noun compounds were originally developed for a 1000-sentence dataset from the British National Corpus that contains 368 two-part noun compounds (Nicholson and Baldwin, 2008). The dataset includes texts from various domains such as literary work, essays, newspaper articles etc. These methods were later adapted to the Wikipedia domain.

Light verb constructions were also identified in the English part of a parallel corpus in which we annotated light verb constructions (14,261 sentence alignment units in size containing 1100 occurrences of light verb constructions). The parallel corpus consists of texts from magazines, novels[1], language books and texts on the European Union are also included. The corpus is available under the Creative Commons license at `http://rgai.inf.u-szeged.hu/mwe`.

3.3 Methodology

We first developed our methods for MWE identification for the source corpora. For both noun compounds and light verb constructions, the corpus that is smaller in size and contains simpler annotation was selected as the source domain. It entails that for noun compounds, the BNC dataset functions as the source domain (containing 1000 sentences and only two-part noun compounds) whereas for light verb constructions, the Wikipedia dataset was selected (containing 4350 sentences and not being annotated for subtypes of light verb constructions).

3.3.1 Detecting noun compounds

For identifying noun compounds in the source domain, we applied the methods 'Match' and 'POS-rules'. Results can be seen in the 'Source' column of Table 2. As it can be expected, POS-rules are beneficial as they improve results.

[1]Not all of the literary texts have been annotated for light verb constructions in the corpus, which made us possible to study the characteristics of the domain and the corpus without having access to the test dataset.

The adaptation process involved the development of more fine-tuned and sophisticated methods considering the domain-specific features of the texts and characteristics of the annotations. Thus, in the case of noun compounds, POS-rules were extended in order to identify noun compounds with more than two parts (e.g. *high school teacher*) because there was no restriction on the length of the annotated noun compounds in Wiki50 and about 20% of them consist of at least 3 parts. The method 'Match' was used as described above. We also implemented a new method for identifying longer noun compounds, which 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'). Finally, we combined the available methods ('Combined').

The **TARGET** column in Table 2 shows results achieved on the target domain when using the original methods whereas the **T+ADAPT** shows those achieved by applying domain-specific methods. The best result can be obtained on the target domain if the three methods are combined, that is, a target-specific method performs best. The process of adaptation is more successful in the case of POS-rules than 'Match', which may be related to the fact that longer units are also identified in Wiki50 and the list we automatically collected from Wikipedia probably contains more noise in the case of longer units. On the other hand, extended POS-rules add to performance.

Another striking fact is that the basic methods (i.e. without any adaptation) perform better on the target domain than on the source domain. The analysis of errors reveals that although it is stated in the BNC paper (Nicholson and Baldwin, 2008) that only sequences of two nouns are annotated, there are in fact longer noun compounds that are also annotated (e.g. *silk jersey halter-neck evening dress*), for which our methods were not prepared. On the other hand, some of the errors are related to annotation errors, for instance, marking noun compounds that contain a proper noun, e.g. *Belfast primary school headmaster*, as simple noun compounds instead of proper nouns (as they should be according to the guidelines), which our system could not identify.

3.3.2 Detecting light verb constructions

Results on the rule-based identification of light verb constructions can be seen in Table 3. In

Method	SOURCE			TARGET			T+ADAPT		
Match	26.93	43.48	33.26	40.45	52.65	45.75	37.7	54.73	44.65
POS-rules	36.91	40.87	38.79	49.04	50.8	49.9	55.56	49.98	52.62
Merge	-	-	-	-	-	-	40.06	57.63	47.26
Combined	-	-	-	-	-	-	59.46	52.48	**55.75**

Table 2: Results of dictionary-based methods for noun compounds in terms of precision, recall and F-measure. SOURCE: source domain, TARGET: target domain without adaptation techniques, T+ADAPT: target domain with adaptation techniques, 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.

the case of the source domain, 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.

Methods developed for the source domain were also evaluated on the target domain without any modification (**TARGET** column). Overall results are lower than those of the source domain, which is especially true for the 'MFV' method: while it performed best on the source domain (41.94%), it considerably declines on the target domain, reaching only 31.18%. The intersection of a verbal and a nominal feature, namely, 'MFV' and 'Stem' yields the best result on the target domain.

Techniques for identifying light verb constructions were also adapted to the other domain. The parallel corpus contained annotation for nominal and participial occurrences of light verb constructions. However, the number of nominal occurrences was negligible (58 out of 1100) hence we aimed at identifying only verbal and participial occurrences in the corpus. For this reason, POS-rules and syntactic rules were extended to treat postmodifiers as well (participial instances of light verb constructions typically occurred as postmodifiers, e.g. *photos taken*).

Since the best method on the Wiki50 corpus (i.e. 'MFV') could not reach such an outstanding result on the parallel corpus, we conducted an analysis of data on the unannotated parts of the parallel corpus. It was revealed that *have* and *go* mostly occurred in non light verb senses in these types of texts. *Have* usually denotes possession as in *have a son* vs. *have a walk* while *go* typically refers to physical movement instead of an abstract change of state (*go home* vs. *go on strike*). The reason for this might be that it is primarily everyday topics that can be found in magazines or novels rather than official or scientific topics, where it is less probable that possession or movement is described. Thus, a new list of typical light verbs was created which did not contain *have* and *go* but included *pay* and *catch* as they seemed to occur quite often in the unannotated parts of the corpus and in this way, an equal number of light verb candidates was used in the different scenarios.

The **T+ADAPT** column of Table 3 shows the results of domain adaptation. As for the individual features, 'MFV' proves to be the most successful on its own, thus, the changes in the verb list are beneficial. Although the features 'Suffix' and 'Stem' were not modified, they perform better after adaptation, which suggests that there might be more deverbal nominal components in the PART class of the target domain. Adaptation techniques add 1.5% to the F-measure on average, however, this value is 6.55% in the case of 'MFV'.

The added value of syntax was also investigated for LVC detection in both the source and the target domains after adaptation. As represented in Table 3, syntax clearly helps in identifying light verb constructions: on average, it adds 2.58% and 2.37% to the F-measure on the source and the target domains, respectively.

4 Discussion

Our adapted methods achieved better results on the target domains than the original ones as regards both noun compounds and light verb constructions. However, the overall results are better for the source domain in the case of light verbs and for the target domain in the case of noun compounds. The latter may be explained by the inconsistent annotation of the BNC dataset – without it, our original methods might have achieved similar results to those on the target domain. As for the former, there is not much difference between

Method	SOURCE			TARGET			T+ADAPT			SOURCE+SYNT			T+ADAPT+SYNT		
POS-rules	7.02	76.63	12.86	5.2	81.47	9.78	5.07	79.4	9.52	9.35	72.55	16.56	6.89	72.97	12.59
Sf	9.62	16.3	12.1	9.7	15.84	12.03	10.5	15.24	12.43	11.52	15.22	13.11	12.81	14.52	13.61
MFV	33.83	55.16	**41.94**	20.59	64.16	31.18	28.81	54.64	37.73	40.21	51.9	**45.31**	34.82	51.19	41.45
St	8.56	50.54	14.64	7.43	62.01	13.26	7.66	61.55	13.62	11.07	47.55	17.96	10.16	56.19	17.2
Sf ∩ MFV	44.05	10.05	16.37	32.13	10.74	16.1	48.31	10.24	16.9	11.42	54.35	18.88	55.03	9.76	16.58
Sf ∪ MFV	19.82	61.41	29.97	15.69	69.26	25.59	19.02	59.64	28.84	23.99	57.88	33.92	23.06	55.95	32.66
Sf ∩ St	10.35	11.14	11.1	10.27	11.41	10.8	11.14	11.07	11.1	12.28	11.14	11.68	14.02	10.59	12.07
Sf ∪ St	8.87	57.61	15.37	7.49	66.44	13.46	7.74	65.71	13.84	11.46	54.35	18.93	10.18	60.12	17.4
MFV ∩ St	39.53	36.96	38.2	27.96	49.4	**35.71**	38.87	43.45	**41.03**	46.55	34.78	39.81	44.04	40.48	**42.18**
MFV ∪ St	10.42	68.75	18.09	7.92	76.78	14.35	8.25	72.74	14.82	13.36	64.67	22.15	10.99	66.9	18.88
Sf ∩ MFV ∩ St	47.37	7.34	12.7	35.09	8.05	13.1	47.41	7.62	13.13	50.0	6.79	11.96	53.98	7.26	12.8
Sf ∪ MFV ∪ St	10.16	72.28	17.82	7.76	78.52	14.13	8.05	74.29	14.53	13.04	68.2	21.89	10.64	68.33	18.49

Table 3: Results of rule-based methods for light verb constructions in terms of precision, recall and F-measure. SOURCE: source domain, TARGET: target domain without adaptation techniques, T+ADAPT: target domain with adaptation techniques, SOURCE+SYNT: source domain with syntactic information, T+ADAPT+SYNT: target domain with adaptation techniques and syntactic information, POS-rules: matching of POS-patterns, Sf: the noun ends in a given suffix, MFV: the verb is among the 12 most frequent light verbs, St: the noun is deverbal.

the performance on the source and the target domains, which might be related to differences in the distribution of (a)typical light verb constructions. However, 'MFV' proves to be the most important feature for both domains, which suggests that with a well-designed domain-specific list of light verb candidates, competitive results can be achieved on any domain, especially if enhanced with syntactic features.

Contrasting the detection of noun compounds and light verb constructions, detecting noun compounds seems to be easier as it achieved better results in terms of F-measure. Indeed, simple features can be successfully applied in identifying noun compounds such as POS-tags and lists because they are syntactically less flexible than light verb constructions on the one hand and a greater part of phrases that match a POS-rule is a noun compound than it is the case for light verb constructions (compare the precision values of the POS-rules method). Thus, the identification of light verb constructions requires morphological, lexical or syntactic features such as the stem of the noun, the lemma of the verb or the dependency relation between the noun and the verb.

The characteristics of the corpora also have an impact on the adaptation process. The smaller the distance between the domains, the easier the adaptation. The topic of the texts were dissimilar in both scenarios (encyclopedia entries in the Wikipedia corpus and miscellaneous topics in the other two corpora) and annotation principles were also quite different in both cases. As our results indicate, the distance is small between the source and the target domain in the case of light verb constructions since similar results can be achieved on the two domains if domain-specific solutions are employed. However, the methods designed for the BNC dataset outperform results on the source domain if evaluated on the target domain, which suggests that the quality of the source data could be improved and thus, no further conclusions can be made on the comparison of the source and target domain in the case of noun compounds.

5 Conclusion

In this paper, we focused on the identification of noun compounds and light verb constructions in different domains, namely, Wikipedia articles and general texts of miscellaneous topics. Our rule-based methods developed for the source domains were adapted to the characteristics of the target domains. Our results indicate that with simple modifications and little effort, our initial methods can be successfully adapted to the target domains as well. For noun compounds, using POS-tagging and lists can lead to acceptable results while a domain-specific list of light verb candidates collected on the basis of sense distribution seems to be essential in detecting light verb constructions.

Obviously, our methods can be further improved. First, the identification of noun compounds relies on an automatically generated list, which can be refined and filtered. Second, stemming of the nominal components of light verb constructions can be enhanced by e.g. wordnet features in order to eliminate false negative matches originating from the stemming principles of the

Porter stemmer (e.g. the stems of *decision* and *decide* do not coincide). Third, the lists of possible light verb candidates can be extended as well. Finally, investigations on other domains and corpora would also be beneficial, which we would like to carry out as future work.

Acknowledgments

This work was supported by the Project "TÁMOP-4.2.1/B-09/1/KONV-2010-0005 – Creating the Center of Excellence at the University of Szeged", supported by the European Union and co-financed by the European Regional Development Fund and by the project BELAMI financed by the National Innovation Office of the Hungarian government.

References

Hassan Al-Haj and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of Coling 2010*, pages 10–18, 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.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. Contrastive filtering of domain-specific multi-word terms from different types of corpora. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 77–80, Beijing, China, August. Coling 2010 Organizing Committee.

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.

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*, MWE '07, pages 41–48, Morristown, NJ, USA. ACL.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL 2003*, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeremy Nicholson and Timothy Baldwin. 2008. Interpreting Compound Nominalisations. In *LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 43–45, Marrakech, Morocco.

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*, pages 57–60, Beijing, China, August.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010b. mwetoolkit: a framework for multiword expression identification. In *Proceedings of LREC'10*, Valletta, Malta, May. ELRA.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010c. Web-based and combined language models: a case study on noun compound identification. In *Coling 2010: Posters*, pages 1041–1049, 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.

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.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of EMNLP-CoNLL 2007*, pages 1034–1043, Prague, Czech Republic, June. ACL.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011a. Detecting noun compounds and light verb constructions: a contrastive study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 116–121, Portland, Oregon, USA, June. ACL.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011b. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.