Domain-Dependent Detection of Light Verb Constructions

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

Abstract

In this paper, we show how our methods developed for identifying light verb constructions can be adapted to different domains and different types of texts. We both experiment with rule-based methods and machine learning approaches. Our results indicate that existing solutions for detecting light verb constructions can be successfully applied to other domains as well and we conclude that even a little amount of annotated target data can notably contribute to performance if a bigger corpus from another domain is also exploited when training.

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; Kim, 2008; Calzolari et al., 2002). They may exhibit peculiar semantic and syntactic features, thus, their NLP treatment is not without problems. Thus, they need to be handled with care in several NLP applications, e.g. in machine translation it must be known that they form one unit hence their parts should not be translated separately. For this, multiword expressions should be identified first.

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 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 domain with both rule-based and machine learning approaches.

The structure of the paper is as follows. First, the characteristics of light verb constructions are presented, then related work is discussed. Our rule-based and machine learning approaches to detecting light verb constructions are presented and our results are analyzed in detail. The paper ends with a conclusion and some ideas on future work are also offered.

2 The Characteristics of Light Verb Constructions

Light verb constructions (LVCs) consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verbal component (also called *light verb*) usually loses its original sense to some extent (e.g. *to take a decision*, *to take sg into consideration*).

In the Wikipedia database used for evaluation (see 4.1) 8.5% of the sentences contain a light verb construction, thus, they are not so frequent in language use. However, they are syntactically flexible: the nominal component and the verb may not be adjacent (in e.g. passive sentences), 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 regarded as one complex predicate.

Light verb constructions deserve special attention in NLP applications for several reasons (Vincze and Csirik, 2010). First, their meaning is not totally compositional, that is, it cannot be computed on the basis of the meanings of the verb and the noun and the way they are related to each other. Thus, the result of translating the parts of the MWE 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 other constructions such as literal verb + noun combinations (e.g. *make a cake*) or idioms (e.g. *make a meal*), thus, their identification cannot be based on solely syntactic patterns. Third, since the syntactic and the semantic head of the construction are not the

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 1–8,
Hissar, Bulgaria, 13 September 2011.

same (the syntactic head being the verb and the semantic head being the noun), they require special treatment when parsing. It can be argued that they form a complex verb similar to phrasal or prepositional verbs.

3 Related Work

Light verb constructions have been paid special attention in NLP literature. Sag et al. (2002) classify them as a subtype of lexicalized phrases and flexible expressions. They are usually distinguished from productive or literal verb + noun constructions on the one hand and idiomatic verb + noun expressions on the other hand: Fazly and Stevenson (2007) use statistical measures in order to classify subtypes of verb + noun combinations.

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: the extracted 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 through the example of identifying English compound nouns in the Genia and Europarl corpora and in general texts (Ramisch et al., 2010b; Ramisch et al., 2010c).

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.

Rule-based domain adaptation techniques are employed in multi-domain named entity recognition as well, and their usability is demonstrated in news stories, broadcast news and informal texts (Chiticariu et al., 2010). They show that domain-specific rules on the classification of ambiguous named entities (e.g. city names as locations or sports clubs) have positive influence on the results.

4 Experiments

For the automatic identification of light verb constructions in corpora, we implemented several rule-based methods and machine learning approaches, which we describe below in detail.

4.1 Corpora Used for Evaluation

We evaluate our approaches on a Wikipedia based corpus, in which several types of multiword expressions (including light verb constructions) and named entities were marked. Two annotators worked on the texts, and 15 articles were annotated by both of them. Differences in annotation were later resolved. As for light verb constructions, the agreement rates between the two annotators were 0.707 (F-measure), 0.698 (Kappa) and 0.5467 (Jaccard), respectively. The corpus contains 368 occurrences of light verb constructions and can be downloaded under the Creative Commons license at `http://www.inf.u-szeged.hu/rgai/mwe`. This dataset proved to be the source domain for the identification of light verb constructions.

Light verb constructions were first identified in the Wikipedia corpus and methods were adapted to 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. In this corpus, different syntactic forms of light verb constructions are annotated:

- verb + noun combinations: *give advice*
- participles: *photos taken*
- nominal forms: *service provider*
- split constructions (i.e. the verb and the noun are not adjacent): *a decision has rarely been made*

The average agreement rate between annotators was 0.7603 (F-measure). The corpus is available under the Creative Commons license at `http://www.inf.u-szeged.hu/rgai/mwe`.

Data on the corpora are shown in Table 1.

4.2 Rule-Based Methods for Identifying Light Verb Constructions

In our investigations, we applied similar methods to those described in Vincze et al. (2011).

[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.

Corpus	Sentence	Token	LVC
Wikipedia	4,350	114,570	368
Parallel	14,262	298,948	1,100

Table 1: Frequency of light verb constructions in different corpora

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 a light verb construction. 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 LVCs.

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 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. The nominal component is typically 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 LVCs. Typically, the syntactic relation between the verb and the nominal component in a light verb construction is `dobj` or `partmod` (using the 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 combined the above methods to identify 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). In order to use the same dataset for evaluating rule based and machine learning methods, we randomly separated the target domain into 70% as training set (used in machine learning approaches) and 30% as test set. As the target domain contained several different topics, we separated all documents into training and test parts. We evaluated our various models in this resulting test set.

4.3 Machine Learning Approaches for Identifying Light Verb Constructions

In addition to the above-described approach, we defined another method for automatically identifying LVCs. The Conditional Random Fields (CRF) classifier was used (MALLET implementations (McCallum, 2002)). The basic feature set includes the following categories (Szarvas et al., 2006):

orthographical features: capitalization, word length, bit information about the word form (contains a digit or not, has uppercase character inside the word, etc.), character level bi/trigrams;

dictionaries of first names, company types, denominators of locations; noun compounds collected from English Wikipedia;

frequency information: frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token which was derived from the Gigaword dataset[2];

shallow linguistic information: part of speech;

contextual information: sentence position, trigger words (the most frequent and unambiguous tokens in a window around the word under investigation) from the training database, the word between quotes, etc.

Some of the above presented LVC specific methods were added to this basic feature set for identifying LVCs. We extended dictionaries with the most frequent verbs like the 'MFV' feature from the rule based methods and a dictionary of the stems of nouns was also added. We extended the orthographical features with the 'Suffix' feature too. As syntax can play a very important role in identifying light verb constructions, we had to extend the shallow linguistic information features with syntactic information.

[2]Linguistic Data Consortium (LDC), catalogId: LDC2003T05

5 Results

We first developed our methods for LVC identification for the source corpus. The Wikipedia dataset is smaller in size and contains simpler annotation, therefore it was selected as the source domain (containing 4350 sentences and not being annotated for subtypes of light verb constructions).

5.1 Rule-Based Approaches

Results on the rule-based identification of light verb constructions can be seen in Table 2. In the case of the source domain, 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 lemma of 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).

Methods developed for the source domain were also evaluated on the target domain without any modification (T w/o ADAPT column). Overall results are lower than in the case 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 24.67%. 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 Wikipedia 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 are 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 2 shows the results of domain adaptation. As for the individual features, 'MFV' proves to be the most successful on its own, thus, the above mentioned changes in the verb list are beneficial. Although the feature 'Suffix' was not modified, it performs better after adaptation, which suggests that there might be more deverbal nominal components with the given endings in the PART class of the target domain, which could not be identified without extended POS-rules. In the light of this, it is perhaps not surprising that its combination with 'MFV' also reaches better results than on the source domain. The intersection of 'MFV' and 'Stem' performs best after adaptation as well. Adaptation techniques add 1.5% to the F-measure on average, however, this value is 6.17% in the case of 'MFV'.

The added value of syntax was also investigated for LVC detection in both the source and the target domains. As represented in Table 3, syntax clearly helps in identifying light verb constructions: on average, it adds 2.58% and 2.45% to the F-measure on the source and the target domains, respectively. On the adapted model, syntactic information adds another 1.39% to performance, thus, adaptation techniques and syntactic features together notably contribute to performance (3.84% on average). The best result on the

Method	SOURCE			T w/o ADAPT			T+ADAPT		
POS-rules	7.02	76.63	12.86	4.28	73.33	8.09	4.28	73.33	8.09
Suffix	9.62	16.3	12.1	9.83	14.58	11.74	9.92	15.42	12.07
MFV	33.83	55.16	**41.94**	16.25	51.25	24.67	22.48	49.17	30.85
Stem	8.56	50.54	14.64	6.55	57.08	11.75	6.55	57.08	11.75
Suffix ∩ MFV	44.05	10.05	16.37	32.35	9.17	14.28	**48.94**	9.58	16.03
Suffix ∪ MFV	19.82	61.41	29.97	13.01	56.67	21.17	15.51	55.0	24.2
Suffix ∩ Stem	10.35	11.14	11.1	11.59	11.25	11.42	11.6	12.08	11.84
Suffix ∪ Stem	8.87	57.61	15.37	6.55	60.42	11.82	6.55	60.42	11.82
MFV ∩ Stem	39.53	36.96	38.2	24.08	40.83	**30.29**	30.72	39.17	**34.43**
MFV ∪ Stem	10.42	68.75	18.09	6.64	67.5	12.09	6.97	67.08	12.63
Suffix ∩ MFV ∩ Stem	**47.37**	7.34	12.7	**40.0**	7.5	12.63	51.35	7.92	13.72
Suffix ∪ MFV ∪ Stem	10.16	**72.28**	17.82	6.53	**69.17**	11.94	6.8	**68.75**	12.39

Table 2: Results of rule-based methods for light verb constructions in terms of precision, recall and F-measure. SOURCE: source domain, T: target domain, ADAPT: adaptation techniques, POS-rules: matching of POS-patterns, Suffix: the noun ends in a given suffix, MFV: the verb is among the 12 most frequent light verbs, Stem: the noun is deverbal.

Corpus	Precision	Recall	F-measure
Wikipedia	60.40	41.85	49.44
Parallel	63.60	39.52	48.75

Table 4: Results of leave-one-out approaches in terms of Precision, Recall and F-measure.

source domain, again, is yielded by the 'MFV' method, which is about 30% above the baseline. On the target domain, it is still the intersection of 'MFV' and 'Stem' that performs best, however, 'MFV' also achieves a good result.

5.2 CRF-Based Approaches

To identify light verb constructions we used the manually annotated corpora (Wikipedia and Parallel) to train CRF classification models (they were evaluated in a leave-one-document-out scheme). Results are shown in Table 4. However, as in the case of the rule-based approach, LVC specific features were adapted to the target corpus. In this way, for instance, the MFV dictionary did not contain *have* and *go* but *pay* and *catch* instead. In the case of the 'Stem' feature, we used domain specific dictionaries. Furthermore, when we trained on the Parallel corpus, we extended the syntax feature rules with `partmod`. On both of the two corpora the CRF based approach can achieve better results than rule-based methods.

For machine learning based domain adaptation we extended our LVC feature set as described in Daumé III (2007). In this way, we extended the above presented basic CRF feature set with domain dependent LVC specific features and with their union. So, some LVC specific features ('MFV' and 'Stem') are represented three times: Wikipedia based, Parallel based and their union, while for the syntax feature, we only used the parallel based one.

As the Wikipedia set was the source domain, we used it as the training set with the above presented extended features, and we extended this training set with randomly selected sentences from the training set of the target domain. We extended the source training set with 10%, 20%, 25%, 33% and 50% of the target domain training sentences in a step-by-step fashion. As Table 5 shows, we evaluated the model trained with the source domain specific feature set (BASE) and the domain adapted trained model (ADAPT) too.

As the results show, the addition of even a little amount of target data has beneficial effects on performance in both the BASE and the ADAPT settings. Obviously, the more target data are available, the better results are achieved. Interestingly, the addition of target data affects precision in a positive way (adding only 10% of parallel data improves precision by about 11%) and recall in a negative way, however, its general effect is that the F-measure improves. Results can be enhanced by applying the domain adapted model. Compared to the base settings, with this feature representation, the F-measure improves 1.515% on average, again primarily due to the higher precision, which

Method	SOURCE+SYNT			T w/o ADAPT + SYNT			T+ADAPT+SYNT		
POS-rules	9.35	72.55	16.56	5.93	69.17	10.92	5.93	69.17	10.92
Suffix	11.52	15.22	13.11	12.1	14.17	13.05	12.1	14.17	13.05
MFV	40.21	51.9	**45.31**	19.54	49.17	27.96	28.28	45.83	34.97
Stem	11.07	47.55	17.96	9.0	54.17	15.43	9.0	54.17	15.43
Suffix ∩ MFV	11.42	54.35	18.88	34.92	9.17	14.52	52.5	8.75	15.0
Suffix ∪ MFV	23.99	57.88	33.92	15.82	54.17	24.48	19.52	51.25	28.27
Suffix ∩ Stem	12.28	11.14	11.68	15.17	11.25	12.92	15.17	11.25	12.92
Suffix ∪ Stem	11.46	54.35	18.93	8.85	57.08	15.32	8.85	57.08	15.32
MFV ∩ Stem	46.55	34.78	39.81	27.81	39.17	**32.53**	37.18	36.25	**36.7**
MFV ∪ Stem	13.36	64.67	22.15	9.0	64.17	15.79	9.56	63.75	16.63
Suffix ∩ MFV ∩ Stem	**50.0**	6.79	11.96	**45.0**	7.5	12.86	**56.67**	7.08	12.59
Suffix ∪ MFV ∪ Stem	13.04	**68.2**	21.89	8.77	**65.42**	15.46	9.21	**65.0**	16.14

Table 3: Results of rule-based methods enhanced by syntactic features for light verb constructions in terms of precision, recall and F-measure. SOURCE: source domain, T: target domain, ADAPT: adaptation techniques, SYNT: syntactic rules, POS-rules: matching of POS-patterns, Suffix: the noun ends in a given suffix, MFV: the verb is among the 12 most frequent light verbs, Stem: the noun is deverbal.

clearly indicates that the domain adaptation techniques applied are optimized for precision in the case of this particular setting and datasets. The advantage of applying both domain-adapted features and adding some target data to the training dataset can be further emphasized if we compare the results achieved without any target data and with the basic feature set (34.88% F-score) and with the 50% of target data added and the adapted feature set (44.65%), thus, an improvement of almost 10% can be observed.

6 Discussion

As the results of the leave-one-out approaches indicate, it is not a trivial task to identify light verb constructions. Sometimes it is very difficult to decide whether an expression is a LVC or not since semantic information is also taken into consideration when defining light verb constructions (i.e. the verb does not totally preserve its original meaning). Furthermore, 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.

For identifying light verb constructions we examined rule-based methods and machine learning based methods too. Rule-based methods were transformed into LVC specific features in machine learning. With the extended feature set the CRF models can achieve better results than the rule-based methods in both corpora.

We also investigated how our rule-based methods and machine learning approaches developed for identifying light verb constructions can be adapted to different domains. For adaptation, characteristics of the corpora must be considered: in our case, the topics of texts determined the modifications in our methods and the implementation of new methods. Our adapted methods achieved better results on the target domains than the original ones in both rule-based and machine learning settings.

The importance of domain-specific annotated data is also underlined by our machine learning experiments. Simple cross-training (i.e. training on Wiki and testing on Parallel) yields relatively poor results but adding some Parallel data to the training dataset efficiently improves results (especially precision).

If rule-based methods and machine learning approaches are contrasted, it can be seen that machine learning settings almost always outperform rule-based methods, the only exception being when there are no Parallel data used in training. Thus suggests that if no annotated target data are available, it might be slightly more fruitful to apply rule-based methods, however, if there are annotated target data and a larger corpus from another domain, domain adaptation techniques and machine learning may be successfully applied. In our settings, even about 1000 annotated sentences from the target domain can considerably improve performance if large outdomain data are also ex-

Method	BASE			ADAPT		
Wiki	29.79	42.08	34.88	31.04	43.33	36.18
Wiki + 10%	40.44	37.91	39.14	42.72	37.91	40.18
Wiki + 20%	40.09	38.75	39.40	43.60	38.33	40.79
Wiki + 25%	41.96	39.16	40.51	47.37	37.5	41.86
Wiki + 33%	45.78	38.41	41.79	46.44	40.83	43.46
Wiki + 50%	47.89	37.91	42.32	49.24	40.83	44.65

Table 5: Results of machine learning approach for light verb constructions in terms of precision, recall and F-measure. BASE: source domain specific feature set trained model, ADAPT: domain adapted trained model.

ploited.

7 Conclusion

In this paper, we focused on the identification of light verb constructions in different domains, namely, Wikipedia articles and general texts of miscellaneous topics. We solved this problem with rule-based methods and machine learning approaches too. Our results show that identifying light verb constructions is a very hard task. Our rule-based methods and results were exploited in the machine learning approaches. We developed our methods for the source domain and then we adapted the characteristics to the target domain. Our results indicate that with simple modifications and little effort, our initial methods can be successfully adapted to the target domain as well. On the other hand, even a little amount of annotated target data can considerably contribute to performance if a bigger corpus from another domain is also exploited when training. As future work, we aim at experimenting on more domains and corpora and we would like to investigate other ways of domain adaptation and machine learning techniques for identifying light verb constructions.

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*, Beijing, China.

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.

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.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of EMNLP 2010*, pages 1002–1012, Stroudsburg, PA, USA. ACL.

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.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16,

Prague, Czech Republic. Association for Computational Linguistics.

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.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

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

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010b. mwetoolkit: a framework for multiword expression identification. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of LREC'10*, Valletta, Malta. 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*, Beijing, China.

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.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In *Discovery Science*, pages 267–278.

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. Association for Computational Lingustics.

Veronika Vincze and János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of Coling 2010*, Beijing, China.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. 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. Association for Computational Linguistics.