Detecting Multiword Expressions and Named Entities in Natural Language Texts

SUMMARY OF PHD THESIS

István Nagy T.

University of Szeged

October 2014

Supervisor:

János Csirik, DSc

Richárd Farkas, PhD

University of Szeged
Doctoral School in Computer Science

1 Introduction

Due to the widespread usage of the sophisticated communication and mobile information devices, the amount of the publicly available information has been increasing at an extraordinary rate. A substantial part of this information is available in textual form, which is written in natural language. The manual processing of this large amount of data requires enormous human effort and financial investments, which can be supported by automatic methods. Natural language processing (NLP) is the study of mathematical and computational modelling of various aspects of natural language and the development of a wide range of computational linguistics systems.

In natural languages there are many ways to express complex human thoughts and ideas. This can be achieved by exploiting compositionality, i.e. concatenating simplex elements of language and thus yielding a more complex meaning that can be computed from the meaning of the original parts and the way they are combined. However, non-compositional phrases can also be found in languages, which are complex phrases that can be decomposed into single meaningful units, but the meaning of the whole phrase cannot (or can only partially) be computed from the meaning of its parts. Such phrases are often called multiword expressions (MWEs) and they display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002). Moreover, MWEs cannot be formulated by directly aggregating the semantics of their constituents. In other words, they are lexical items that contain space. Therefore, for the natural language processing applications that require the semantic processing of texts, the detection of multiword expressions is an indispensable part.

In this thesis, we will focus on the automatic detection of English and Hungarian multiword expressions. As MWEs occur quite frequently in both languages and their proper treatment of MWEs is essential for several natural language processing applications like information extraction and retrieval, terminology extraction and machine translation, it is important to identify multiword expressions in context. For example, in machine translation we must know that MWEs form one semantic unit, hence their parts should not be translated separately. For this, multiword expressions should be identified first in the text to be translated.

Here, we will present different types of multiwords expressions which we will concentrate on in this thesis.

Nominal compounds (NCs) are subtypes of multiword expressions. NCs are lexical units that consist of two or more elements that exist on their own, the unit functions as a noun and it usually has some extra meaning component compared with the meanings of the original parts, such as the following English and Hungarian examples:

(a) *black sheep – fekete bárány*

(b) *stock car – marhavagon*

Light verb constructions (LVCs) are another subtype of MWEs. LVCs are verb and noun combinations in which the verb has lost its meaning to some extent and the noun is used in one of its original senses. Examples taken from English and Hungarian are shown in the following:

(a) English:
to take measure
to play a role

(b) Hungarian:
őrizetbe vesz "to take into custody"
döntést hoz "to take a decision"

Named entities (NEs) are another class of linguistic elements that require specific treatment in many NLP applications ranging from information retrieval to machine translation. A named entity is a phrase in the text which uniquely refers to an entity of the world, like the name of an organization or location. Named entities often consist of more than one word; that is, they can be viewed as a specific type of multiword expressions / nominal compounds (Jackendoff, 1997; Sag et al., 2002). Similar to multiword expressions, the meaning of multiword named entities cannot be traced back to their parts. For instance, *Ford Focus* refers to a car and has nothing to do with the original meaning of *ford* or *focus*; thus it is justifiable to treat the whole expression as one unit. NEs function as nouns just like NCs. Moreover, the linguistic relatedness is demonstrated by the fact that an NC may include an NE (***FBI** special agent*) and may be part of an NE (*Tallulah **High School***), and an

NE may contain another NE (as *Oxford* and *Oxford University* in *Oxford University Press*). On the other hand, sometimes it is not unequivocal to decide whether a multiword unit is a nominal compound or a named entity (e.g. *Attorney General*). Although both nominal compounds and multiword named entities consist of more than one word, they form one semantic unit and thus, they should be treated as one unit in NLP systems. Moreover, as they behave similarly we will argue that the same methods should be applied for their automatic detection.

The main aim of this thesis is to automatically identify different types of multiword expressions in English and Hungarian raw texts. As verbal MWEs and multiword named entities are quite frequent in both languages, we seek to identify them together with English nominal compounds, and we will implement several machine learning-based approaches for these tasks.

2 Results of the Thesis

The main results achieved in this thesis will be summarized in the next sections, and the papers in which these results have been published are also listed, together with the author's main contributions.

2.1 English Nominal Compound Detection with Wikipedia-Based Methods

We examined dictionary and machine learning-based methods for identifying noun compounds on different corpora. These approaches made intensive use of Wikipedia data. In order to automatically identify nominal compounds, we also applied a machine learning-based method. We also illustrated how previously identified nominal compounds affect named entity recognition and vice versa, how nominal compound detection is supported by identified named entities. We found that previous knowledge of nominal compounds can enhance NER, while previously identified NEs can assist the nominal compound identification process. We also looked at the effectiveness of the machine learning-based method when it was trained on an automatically generated silver standard corpus and we demonstrated that this approach can also provide acceptable results. Moreover, we also investigated how the size of an automatically generated silver standard corpus can affect the performance of our machine learning based method. The results we obtained demonstrate that the bigger the dataset, the better the performance will be **(Thesis 1)**.

The main results include:

- We presented results on NC detection got by applying our Wikipedia-based dictionary labeling method on the Wiki50 corpus and BNC dataset.
- We presented results for NC detection after applying of our supervised machine learning-based model on the Wiki50 corpus. This approach achieved the highest F-score value as it used a supervised model.
- We also investigated how the size of an automatically generated silver standard corpus could affect the performance of our machine learning-based method. The results we obtained demonstrate that the bigger the dataset, the better the performance will be.
- We presented the results of our experiments on how the size of Wikipedia could improve the performance of our Wikipedia-based dictionary labeling method for detecting nominal compounds. We found that the growth of Wikipedia improved the performance, especially the recall score, but the rate of improvement decreased over time.
- We investigated the usefulness of NCs in named entity recognition and vice versa, and examined how NC detection was supported by identified named entities. The results indicated that the knowledge of named entities is useful in the NC identification process and known nominal compounds can assist named entity recognition.

In Vincze et al. (2011b), the Wiki50 corpus was presented along with the primary results got by using dictionary lookup methods. The author developed the dictionary-based method to automatically detect nominal compounds. One of the co-authors annotated the corpus and provided the linguistic background.

In Vincze et al. (2011a), nominal compounds are identified in running text with rule-based methods. The author developed the dictionary lookup and rule-based methods for the automatic detection of nominal compounds and light verb constructions, and compared the effect of the different features. The co-authors were responsible for linguistic analysis of the data.

In Nagy T. et al. (2011), nominal compounds and named entities were identified, and we investigated how they can contribute to keyphrase extraction, furthermore we also examined how previously identified nominal compounds affected Named Entity Recognition and vice versa, how nominal compound detection is supported by identified named entities. The author implemented the machine learning-based nominal compound detector and examined the effectiveness of the previously known named entities and nominal compounds on Named Entity Recognition and nominal compound detection, respectively. The co-authors were responsible for the linguistic analysis of nominal compounds and named entities and keyphrase extraction experiments.

In Nagy and Vincze (2013), Wikipedia-based methods were presented for the automatical detection of nominal compounds. The author investigated how the size of an automatically generated silver standard corpus can affect the performance of the machine learning-based method, as well as how the growth of the Wikipedia added to the performance of the dictionary lookup method. The co-author was responsible for the linguistic background.

2.2 Named Entity Recognition Problems in Web Mining

We consider named entities similar to nominal compounds as they form one semantic unit and consist of more than one word and they function as a noun. Therefore a similar approach could be applied for their recognition as in the case of nominal compounds. There are several NER methods described, but we focused on Web Mining-based named entity recognition problems, such as Researcher Affiliation Extraction, Person Attribute Extraction and Company Contact Information Extraction, which make use of named entity recognition **(Thesis 2)**.

The main results include:

- Here, we presented three different Named Entity Recognition problems from the area of web mining, in two different languages, namely English and Hungarian.

- As we found that most of the useful information was available in natural text format in webpages, **we focused on the raw textual parts of the webpages** instead of the structured parts.

- As only a small portion of extracted textual paragraphs contained useful information, we **developed attribute-specific relevant section selection modules**. Our filtering method exploited the paragraphs containing a current attribute (`positive paragraphs`).

- We treated **named entities similar to nominal compounds as they form one semantic unit**, consist of more than one word and function as a noun. Also, we found in three NER datasets that the majority part of named entities were multiword named entities. Therefore similar machine learning-based methods could be applied just like in the case of nominal compounds.
- We were able to extract attributes belonging to the same semantic class in a better way via machine learning approaches when **we placed the attribute classes into logical groups** and we assumed ordered relations among the coherent attributes.
- We presented a Web Content Mining system for **gathering affiliation information** from the homepages of researchers.
- Our attribute extraction method efficiently **extracted the different types of person-related attributes** from webpages and we achieved top results in the WePS3 challenge.
- We presented our approaches to detect **names and addresses of companies with rule-based and machine learning-based methods** on the webpages of companies.

In Nagy et al. (2009), information about researchers' affiliations is identified from webpages. The problem of person attribute extraction from webpages is described in Nagy T. (2012). The author participated in the third WePS challenge (Artiles et al., 2010) and achieved top results on the person attribute extraction subtask. The extraction of company contact information is presented in Nagy T. (2009).

2.3 Sequence Labeling for Detecting English and Hungarian Light Verb Constructions

We implemented our conditional random fields based tool for identifying verbal light verb constructions in running texts. The flexibility of the tool is demonstrated on two, typologically different languages, namely, English and Hungarian. Furthermore, different types of texts may contain different types of light verb constructions, and the frequency of light verb constructions may differ from domain to domain. Hence we focused on the portability of models trained on different corpora and we also investigated the effect of simple domain adaptation techniques to attempt to reduce the gap between the domains. Our results show that in spite of their special domain characteristics, out-domain data can also contribute to successful LVC detection in different domains **(Thesis 3)**.

The main results include:

- Here, we addressed a **broader range** of LVCs than previous studies did. In contrast to most of them, we did not just focus on verb-object pairs. Instead, we identified LVCs that contained adpositional complements or nouns in an oblique case.
- We introduced our **conditional random fields** based state-of-the-art **tool** for detecting LVCs, which makes use of contextual (shallow linguistic) features and is able to produce satisfactory results for all of the domains and languages used.
- We reported our results for Hungarian and English corpora as well, which allowed us to draw some conclusions on the **multilingual aspects** of LVC detection.
- In our experiments, we made use of three corpora for both languages. The corpora belong to different domains, namely short news, law and newspaper texts. This selection of data made it possible for us to compare the **domain-specific characteristics** of LVC detection in both languages. We reported results for three domains in two languages, and this allowed us to make **cross-lingual comparisons** for each domain.
- We applied **domain adaptation** techniques in order to reduce the distance between domains in a setting where only limited annotated datasets are available for one of the domains.

In Vincze et al. (2013b), verbal light verb constructions were identified by using a conditional random fields-based tool. The author implemented the machine learning-based method on English and Hungarian, furthermore he applied domain adaptation techniques. He also investigated the effect of simple domain adaptation techniques to reduce the gap between any two domains. The co-authors of the paper were responsible for the linguistic background and the statistical analysis of the corpus data.

2.4 Full-coverage Identification of English and Hungarian Light Verb Constructions

The CRF-based model could automatically detect English and Hungarian verbal LVCs, but this approach could not handle other types of LVCs like SPLIT and PART. Therefore, we focused on the full-coverage identification of light verb constructions. Our offered approach first syntactically parse each sentence and extracted potential LVCs with different candidate extraction methods. Moreover, we also investigated the performance of different candidate extraction methods on full-coverage LVC annotated corpora on English and Hungarian, where we found that less severe candidate extraction methods should be applied. Then we followed a machine learning approach that makes use of an extended and rich feature set to select LVCs among extracted candidates **(Thesis 4)**.

The main results include:

- We introduced and evaluated systems for **identifying all LVCs and all individual LVC occurrences** in English and Hungarian running texts and we did not restrict ourselves to certain specific types of LVCs.
- We systematically **compared and evaluated different candidate extraction methods** (earlier published methods and new solutions implemented by us).
- We defined and evaluated several **new feature templates** like semantic or morphological features to select LVCs in context from extracted candidates. For each of the two languages, each type of feature contributed to the overall performance.
- We applied both **language independent and language specific features**: we compared whether the same set of features could be used for both languages, then investigated the benefits of integrating language specific features into the systems and we explored how the systems could be further improved.
- The method proved to be sufficiently robust as it achieved **approximately the same scores on two typologically different languages**.

In Nagy T. et al. (2013), a system was introduced that enables the full coverage identification of English LVCs in running texts. The author implemented the machine learning

			Thesis			
			1	2	3	4
RANLP	2011	(Nagy T. et al., 2011)	•			
RANLP	2011	(Vincze et al., 2011b)	•			
MWEWS	2011	(Vincze et al., 2011a)	•			
TSD	2013	(Nagy and Vincze, 2013)	•			
ACTA	2012	(Nagy T., 2012)		•		
NLPIR4DL	2009	(Nagy et al., 2009)		•		
OTDK	2009	(Nagy T., 2009)		•		
ACM	2013	(Vincze et al., 2013b)			•	
ACL	2013	(Vincze et al., 2013a)				•
IJCNLP	2013	(Nagy T. et al., 2013)				•

Table 1: The relation between the theses and the corresponding publications.

based method, he added some new features and developed syntax-based candidate extraction methods, however, experimental results are treated as a shared contribution of all authors. The co-authors were responsible for the linguistic background and the idea of the full-coverage identification of LVCs. In Vincze et al. (2013a), Hungarian and English LVCs were identified in free texts. The author contrasted the performance of the applied methods and applied language-specific features on these typologically different languages. The co-authors were responsible for the linguistic background and the interlingual comparisons.

The relationship of the publications and the above listed theses is visually represented in Table 1.

3 Conclusions and Future Work

In this thesis, we focused on the automatic detection of multiword expressions in natural language texts. On the basis of the main contributions, we can argue that:

- supervised machine learning methods can be successfully applied for the automatic detection of different types of multiword expressions in natural language texts;
- machine learning-based multiword expression detection can be successfully carried out for English as well as for Hungarian;
- Web Mining-based named entity recognition problems required a similar approach as in the case of nominal compounds;

- the previous knowledge of nominal compounds can enhance NER, while previously identified NEs can assist the nominal compound identification process;
- our machine learning-based method can also provide acceptable results when it was trained on an automatically generated silver standard corpus;
- our conditional random fields based tool can be successfully applied for identifying verbal light verb constructions in two, typologically different languages, namely, English and Hungarian;
- there are domain specificities of multiword expression distribution;
- a small change in the domain generally requires new manually labeled training corpus. Hence, domain adaptation techniques may help diminish the distance between domains in multiword expressions detection;
- our syntax-based method can be successfully applied for the full-coverage identification of light verb constructions. As a first step, a data-driven candidate extraction method can be utilized.

Besides the main points described above, the results of the thesis may be applicable in other fields of NLP research as well as in other disciplines. In several natural language processing applications like information extraction and retrieval, terminology extraction, machine translation and document classification, it is important to identify multiword expressions in context. For example, in machine translation we must know that MWEs form one semantic unit, hence their parts should not be translated separately. For this, MWEs should be identified first in the text to be translated. Information retrieval may also be enhanced by detecting multiword expressions.

In another example, LVCs denote one event and again they should be treated as one unit in event extraction tasks. As before, the extraction of events must be preceded by identifying LVCs in the text.

In the future, we would like to improve our systems by conducting a detailed analysis of the effect of the features included. Later, we also plan to adapt our tools to other types of multiword expressions, like verb particle–constructions and conduct further experiments on languages other than English and Hungarian. Moreover, we can improve the applied methods in each language and each type of MWEs by implementing other language-specific features as well. Moreover, we also would like to provide a standardized (i.e. language-independent) representation of different types of multiword expressions that can be used in machine learning experiments in a language-independent context.

We believe that our research on the automatic detection of multiword expressions can be successfully exploited in several NLP tasks and it will contribute to develop novel approaches in many fields of natural language processing.

References

Artiles, Javier; Borthwick, Andrew; Gonzalo, Julio; Sekine, Satoshi; Amigó, Enrique. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF).*

Calzolari, Nicoletta; Fillmore, Charles; Grishman, Ralph; Ide, Nancy; Lenci, Alessandro; MacLeod, Catherine; Zampolli, Antonio. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 1934–1940, Las Palmas.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

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

Nagy, István; Vincze, Veronika. 2013. English Nominal Compound Detection with Wikipedia-Based Methods. In Matousek, Václav; Mautner, Pavel; Pavelka, Tomás (eds.), *Proceedings of the 16th International Conference on Text, Speech and Dialogue, TSD 2013*, Lecture Notes in Computer Science, pp. 225–232. Springer, Berlin / Heidelberg, September.

Nagy, István; Farkas, Richárd; Jelasity, Márk. 2009. Researcher affiliation extraction from homepages. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pp. 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nagy T., István; Berend, Gábor; Vincze, Veronika. 2011. Noun compound and named entity recognition and their usability in keyphrase extraction. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Nagy T., István; Vincze, Veronika; Farkas, Richárd. 2013. Full-coverage Identification of English Light Verb Constructions. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 329–337, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Nagy T., István. 2009. Összetett rendszer vállalkozások címeinek webről történő automatikus összegyűjtésére [Complex system for automatic detection of addresses of companies from Web]. In *XXIX. Országos Tudományos Diákköri Konferencia OTDK Informatikai szekció*. Debrecen.

Nagy T., István. 2012. Person attribute extraction from the textual parts of web pages. *Acta Cybernetica*, 20(3):419–440.

Sag, Ivan A.; Baldwin, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pp. 1–15, Mexico City, Mexico.

Vincze, Veronika; Nagy T., István; Berend, Gábor. 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*, pp. 116–121, Portland, Oregon, USA, June. ACL.

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

Vincze, Veronika; Nagy T., István; Farkas, Richárd. 2013a. Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 255–261, Sofia, Bulgaria, August. Association for Computational Linguistics.

Vincze, Veronika; Nagy T., István; Zsibrita, János. 2013b. Learning to detect English and Hungarian light verb constructions. *ACM Trans. Speech Lang. Process.*, 10(2):6:1–6:25, June.