Acquisition of domain-specific multiword expressions in Serbian

Vesna Pajić*, Miloš Pajić*

* Center for data mining and bioinformatics, University of Belgrade - Faculty of Agriculture, Nemanjina 6, 11080 Belgrade, Serbia

Introduction

The term “multiword expression” (MWE) denotes linguistic expressions composed of two or more words functioning as a single unit at semantic level (Calzolari et al., 2002). Their processing is a major challenge for computer science, due to their non-compositionality (the meaning of the expression cannot be determine from the meanings of its components), as one of the main characteristics of MWE. On the other hand, MWEs are very frequent in both written and spoken language (over 50% of the lexical units used, according to Ramisch (2009)).

As such, they have been studied intensively for the last two decades in a number of NLP based researches for several languages (Laporte and Voyatzi, 2008; Moszczynski, 2007; Nakov and Hearst, 2013; Ramisch et al., 2008; Seretan and Wehrli, 2013;Villavicencio et al. 2007), but the work on MWEs in English still dominates.

The morphological dictionary of multi-word units for Serbian is in DELAC format and has at present 11,324 entries. Its development is a very challenging task, due to the rich morphological system of Serbian. (Krstev et al., 2010) automated the production of dictionary lemmas for a given list of MWEs, using e-dictionaries of Serbian simple words (Krstev et al., 2013). Here, we aim to automate the process of identifying MWEs in a text, in order to further extend the existing resources for MWE processing in Serbian.

Domain-specific MWE

When seen in the context of domain-specific texts, MWEs constitute a significant portion of terminology; over 70% of the terms are complex lexical units, composed of more than one word (Krieger and Finatto (2004)). Domain-specific MWEs are difficult to detect automatically using the existing methods, since they often have relatively few instances in big corpora and therefore cannot be spotted by exploiting their statistical properties. To improve the extraction of MWE in general, different authors suggest some kind of morphosyntactic analysis (Rayson et al., 2009). When it comes to domain-specific MWEs, it is necessary to know the relevant syntactic structures used in particular language and the domain for expressing specific concepts.

Experiments in Serbian

The main objective of the research was to populate the existing morphological dictionary of compound lexical units in Serbian with multiword terms from the agricultural domain, such as: ***silažni kukuruz***, *corn silage* (eng); ***semenski krompir***, *seed potato* (eng), ***organska proizvodnja***, *organic farming*(eng.); ***precizna poljoprivreda***, *precision agriculture* (eng); ***obrada teških zemljišta***, *treatment of heavy soils* (lit.); ***žitni kombajn***, *weed harvester* (lit.); etc. (Krstev et al, 2011) already suggested a method for automatic extraction of MWEs from literal text in Serbian. Here, we will try to determine what results could be achieved in a limited domain text.

The text collection used for preliminary analysis consists of scientific papers from the domain of agricultural engineering. We analyzed it with Unitex 3.0[1], a software tool for linguistically based processing of texts. The corpus has over 716,000 tokens, out of which 94,572 are recognized as simple word forms, 8,340 as compound lexical entries (already included in morphological Serbian dictionary) and 5,330 as unknown simple words (words

[1] *http://www-igm.univ-mlv.fr/~unitex/*

not listed in the dictionary). Among compound entries, only 1,540 are multiword expressions in different word forms, while the rest are multiword numerals.

In order to extract new multiword terms from the text, we lemmatized it first to avoid recognition of different morphemes of same lexical units as different MWEs. Then, the text was searched for expressions that have the structure <A><N>. In order to distinguish MWEs from non-MWEs of the same syntactic form, we used frequencies in the corpus. Only expressions with high frequency were taken as multiword terms.

There were 29,753 expressions having the syntactic structure <A><N>, some of them having the same lemma, but occurring in several forms. We estimate that, based on the obtained results, over 50% of the MWEs found in this way are actually multiword terms. Terms representing some of the most important agricultural concepts are extracted, such as ***angažovana snaga***, *engaged power* (eng.); ***aromatično bilje***, *aromatic plants* (eng.); ***genetički potencijal***, *genetic potential* (eng.); ***hidraulički sistem***, *hydraulic system* (eng.) etc.

Discussion

The creation of lexical resources for Serbian language is of the most importance for NLP community in Serbia. Lexicons of MWEs for Serbian are still in the initial phase of development, and different acquisition techniques are needed. Having a rich morphological system, Serbian can not be processed with the same, sometimes neither with similar methods as English. The work presented here aims to contribute significantly in the process of the resources creation for Serbian. In our experiment, we searched only for one possible syntactic structure of multiword terms. We plan to conduct researches with other structures in the same manner, such as <N><A><N> and others. The overall objective is to develop methods that can be used in general, for acquisition of terms from different domains.

References

Calzolari, N, Fillmore, C., Grishman, R, Ide, N, Lenci, A., MacLeod, C., Zampolli. A (2002). Towards best practicefor multiword expressions in computational lexicons. In Proceedings of LREC 2002, pages 1934–1940, Las Palmas.

Krieger, M., Finatto, M. J. B. (2004). Introdução à Terminologia: teoria & prática. Editora Contexto, São Paulo, SP, Brazil. 223 p.

Krstev, C., Stanković, R., Obradović, I., Vitas, D., Utvić, M. (2010) Automatic Construction of a Morphological Dictionary of Multi-Word Units. In proceeding of: Advances in Natural Language Processing, 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010

Laporte, É. and Voyatzi, S. (2008). An electronic dictionary of French multiword adverbs. In 2007. Proceedings of the ACL workshop: A broader perspective on multiword expressions., pages 31–34.

Moszczynski, R. (2007). A practical classification of multiword expressions. In Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, 19-24.

Nakov, P., Hearst, M.A. (2013) Semantic interpretation of noun compounds using verbal and other paraphrases. TSLP 10(3): 13

Ramisch, C., Schreiner, P., Idiart, M., Villavicencio A. (2008) An evaluation of methods for the extraction of multiword expressions. In Proceedings of the ACL workshop: A broader perspective on multiword expressions, pages 50–53.

Ramisch, C. (2009). Multiword terminology extraction for domain-specific documents. Master's thesis, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, France.79p.

Rayson, P. S. Piao, S. Sharoff, S. Evert, B. Villada Moiron. (2009) Multiword expressions: hard going or plain sailing? Journal of Language Resources and Evaluation.

Seretan, V., Wehrli, E. (2013): Syntactic concordancing and multi-word expression detection. IJDMMM 5(2): 158-181

Villavicencio,A., Bond, F., Korhonen, A., McCarthy, D. (2005). Introduction to the special issue on multiword expressions: having a crack at a hard nut. Journal of Computer Speech and Language Processing, 19(4):365–377.

Krstev, C., Obradović, I., Stanković, R., Vitas, D. (2013) An Approach to Efficient Processing of Multi-word Units, in Computational Linguistics - Applications, pp. 109-229, 2013.

Krstev, C., Vitas, D., Trtovac, A. (2011) Orwell's 1984 – the Case of Serbian Revisited, in Proceedings of 5th Language & Technology Conference, November 25-27, 2011, Poznań, Poland.