Multiword Term Extraction: the Multilingual Perspective

Vesna Pajić[1], Cvetana Krstev[2], Ranka Stanković[3] and Staša Vujičić Stanković[4]

[1] *University of Belgrade, Faculty of Agriculture, Nemanjina 6, Zemun, Belgrade*
[2] *University of Belgrade, Faculty of Philology, Studentski trg 3, Belgrade*
[3] *University of Belgrade, Faculty of Mining and Geology, Đušina 7, Belgrade*
[4] *University of Belgrade, Faculty of Mathematics, Studentski trg 16, Belgrade*

1. Introduction

A multiword expression (MWE) is a linguistic construction of two or more words whose semantic and/or syntactic property cannot fully be predicted from those of its components, and thus is functioning as a single unit (Calzolari et al., 2002). According to Manning and Schütze (1999), MWEs have three important characteristics: non-compositionality (semantically opaque), non-modifiability (syntactically rigid) and non-substitutability (lexically determined). A term is a lexical unit that has an unambiguous meaning when used in a text of a specific domain, representing a concept of a knowledge area. Terms can be single words or MWEs. Here, we are interested in domain-specific terms that are composed of more than one word, i.e. Multi-Word Terms (MWTs). MWTs constitute a significant portion of terminology lexicons; over 70% of the terms are complex lexical units, composed of two or more words (Krieger and Finatto (2004)).

MWTs have additional important characteristic, beside the three already mentioned. It is their domain-dependency. One word combination can be an MWT in one domain, but when seen in general context or in some other domain, it need not be an MWE at all. For example, the expression "*sistem obrade*" (eng. *tillage system,* lit. *system of processing*) is an MWT for agricultural domain, but in general corpora it would not be recognized as an MWE. Because of this characteristic, MWT processing has to be studied separately from processing general MWEs.

MWT lexicons are very valuable for NLP processing tasks, but their production is time-consuming. Methods for MWT automatic acquisition have been developed, in order to help the creation of resources. Here, we compare these methods, and investigate their transitivity to languages other than those they were originally designed for. Additionally, we explore the possibilities of using existing multilingual resources (parallel corpora and multilingual thesauri) in term extraction processes and how they can be used for the creation of new resources.

2. Multiword Term Extraction

2.1. Existing approaches to MWT extraction

Methods developed for MWTs extraction usually integrate statistical information, such as frequencies of n-grams and collocations, with lexical information, such as syntactic patterns of expressions (Ramisch et al., 2008). One of the earlier researches is the work of Justeson and Katz (1995). They used a set of syntactic patterns for MWT extraction, combined with raw frequencies of term candidates for filtering purposes. After that, different authors tried to improve precision and recall by using different measurements of termhood, such as T-score, the log-likelihood ratio (LLR), or the C/NC value, instead of raw frequencies and by adapting the syntactic patterns to particular languages (Boguraev and Kennedy 1999; Maynard and Ananiadou 2000; Park et al. 2002; Koeva 2007; Sclano and Velardi 2007; Vivaldi and Rodriguez 2007; Savary and Zaborowski, 2012). Here, we will give a detailed review of related work and the methods and measurements that are widely used.

2.2. Comparison of MWT extraction for different languages

The architecture of MWT extraction systems, consisting of syntactic analysis part and statistically based filtering, can be seen as language-independent. However, the parts themselves depend on a language. Methods developed for one language cannot be used for another one without the loss of precision and/or recall. In some cases, the same algorithms can be improved with additional modifications, but this is not always the case. The comparison of algorithm efficiency when applied to different group of languages will be given, emphasizing the case of morphologically complex languages.

3. Multilingual aspect of MWTs

The majority of methods for MWT extraction were initially developed for English. As in other fields of computational linguistics, MWT extraction for many languages lacks research, methods and resources. There are several important questions regarding multilingual aspect of MWT: (i) is it possible to make a transition of a MWT extraction method from one language to another one, (ii) if so, to what extent, (iii) what modifications need to be done? In order to answer these questions, we will explore different issues of MWTs in different languages and compare them. The main work will be done in comparing different syntactic patterns in one language with the patterns of corresponding terms in another one.

Moreover, it would be beneficial to see whether existing multilingual resources can be used and in what part of the extraction process. By multilingual resources, we consider any electronic linguistic resource that contains information on some lexical or textual unit in more than one language. These can be aligned corpora, used for terms extraction, or multilingual thesauri, lexicons and dictionaries, used in or created by term extraction process. Some multilingual aspects of terminology and MWEs are already discussed in (Krstev et al, 2009; Stanković et al, 2010; Stanković et al, 2012; Vitas and Krstev, 2012). Here, we give a review of existing resources, as well as of research that utilize this kind of resources in the term extraction process.

4. Experiments with aligned corpora and multilingual resources

We experiment with two kinds of resources: aligned corpora and multilingual thesauri. In the first experiment, we automatically produce a multilingual terminological lexicon from the aligned corpora. The corpora consist of scientific papers published in Serbia bilingually (in Serbian and in English) belonging to several domains: library and information sciences, architecture and urbanism, and mining engineering.

In the second experiment, we automatically enrich existing electronic dictionary from other resources, such as Agrovoc, Eurovoc and alike. For example, there are 6,611 terms in Serbian in Eurovoc, with 5,091 of them being MWTs. This kind of resource can significantly contribute to Serbian electronic dictionary. Moreover, although here applied to terms in Serbian, the same approach can be used for any other language.

5. Conclusion and Future work

Domain-specific texts represent a large knowledge source expressed in natural language and are therefore of great interest for NLP. On the other hand, their linguistic properties differ from general-purpose language. Processing MWTs, as one of the main factors of a domain's terminology, is maybe the most important task in many different NLP tasks, such as information retrieval, text classification, machine translation, natural language generation, multilingual text search etc. Multilingual resources, methods and approaches can significantly improve processing of MWT.

In this work, we give a comprehensive overview of state-of-the-art in MWT extraction. Moreover, we suggest some directives how to choose a method or a resources type

for MWT extraction, depending on properties of particular language. The experiments show how the creation of MWT lexicons can be significantly improved by using multilingual resources. We plan to continue with developing resources for different domains.

References

Boguraev, B., and Kennedy, C. (1999) Applications of term identification technology: domain description and content characterisation. Natural Language Engineering 5(01): 17-44.

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.

Justeson, J. and Katz, S. (1995) Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1):9–27, 1995.

Koeva, S. (2007) Multi-Word Term Extraction for Bulgarian. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, Prague, Czech Republic, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 59-66.

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

Krstev, C., Stanković, R., Vitas, D. and Koeva, S. (2009) "E-Connecting Balkan Languages", in Proceedings of the Workshop Workshop on Multilingual resources, technologies and evaluation for Central and Eastern European Languages, 17 September 2009, Borovets, Bulgaria, eds. C. Vertan, S. Piperidis, E. Paskaleva and Milena Slavcheva, pp. 23-29, 2009. ISBN 978-954-452-008-3.

Manning, C. and Schütze, H. (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, USA, 1999. ISBN 0-262-13360-1.

Maynard, D., and Ananiadou, S. (2000) Identifying terms by their family and friends. In Proceedings of the 18th Conference on Computational Linguistics- Volume 1, Saarbrucken, Germany, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 530-536.

Park, Y., Byrd, R., and Boguraev, B. (2002) Automatic Glossary Extraction: Beyond Terminology Identification. In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, Saarbrucken, Germany, Stroudsburg, PA, USA: As- sociation for Computational Linguistics, pp. 1-7.

Savary, A., and Zaborowski, B. (2012) SEJFEK - a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In 24th International Conference on Computational Linguistics, Mumbai, India, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 195.

Sclano, F., and Velardi, P. (2007) TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In Enterprise Interoperability II, Springer, Berlin: Springer Verlag, pp. 287-290.

Stanković, R., Obradović, I. and Kitanović, O. (2010), GIS Application Improvement with Multilingual Lexical and Terminological Resources, Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2010, Valetta, Malta, May 2010, European Language Resources Association Valetta, Malta, 2010, pp. 2283-2287, 2-9517408-6-7, http://www.lrec-conf.org/proceedings/lrec2010/pdf/57_Paper.pdf

Stanković, R., Krstev, C., Obradović, I., Trtovac, A. and Utvić, M. (2012) A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals, Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, May 2012, Istanbul, Turkey, N. Calzolari et al. (eds.), European Language Resources Association Istanbul, Turkey, 2012, pp. 1710-1717, ISBN 978-2-9517408-7-7, http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf.

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

Vivaldi, J., and Rodriguez, H. (2007) Evaluation of terms and term extraction systems: A practical approach. Terminology 13(2):225-248.

Vitas, D. and Krstev, C. (2012) Construction and Exploitation of X-Serbian Bitexts. In Cristina Vertan and Walther v.Hahn (eds.) Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation, pp. 207-227, Cambridge Scholars Publishing, 2012. ISBN (13) 978-1-4438-3878-8.