Short Term Scientific Mission- REPORT

KEYSTONE COST Action IC1302

STSM Applicant:	Olivera Kitanović, Ph.D.
Home Institution:	Faculty of Mining and Geology, University of Belgrade, Serbia
Host Institution:	Department of Informatics, University of Rijeka, Croatia
Duration:	14th-30th November, 2016

STSM title: Cross-lingual extraction and aligning of keywords and key-phrases from monolingual and multilingual textual resources

1. Purpose of the STSM

The purpose of the STSM was research collaboration in the domain of the extraction of keywords and key-phrases from documents in Serbian, Croatian and English language, based on the Selectivity-Based Keyword Extraction (SBKE) method. Selectivity-based extraction does not require linguistic knowledge as it is derived purely from statistical and structural information of the network. The main focus of our research was to improve the results of the SBKE method using stemmers and lemmatization. This research was inspired by the KEYSTONE Conference paper under title "Network-Enabled Keyword Extraction for Under-Resourced Languages" [4].

Furthermore, our goal was to establish a new collaboration between the LangNet team from the Department of Informatics, University of Rijeka and the Human Language Technology Group from University of Belgrade.

2. Description of the work carried out during the STSM

During my visit to the Department of Informatics, University of Rijeka in the period of 17-30 November, I had the opportunity to work with professor Sanda Martinčić - Ipšić and Slobodan Beliga, who shared with me their expertise in keywords extraction based on Selectivity-Based Keyword Extraction (SBKE) method [1,2].

SBKE method is based on keyword extraction from the complex network. Node selectivity measure is defined as the average weight distribution on the links of a single node and used in procedure of keyword candidate extraction [4]. In the directed network, the in/out strength s_i of the node *i* is defined as the number of its incoming and outgoing links, that is:

$$s_i^{in/out} = \sum_j w_{ji/ij}.$$

The selectivity measure is introduced in [9]. It is actually an average strength of a node. For the node *i* the selectivity is calculated as a fraction of the node weight and node degree:

$$e_i = \frac{s_i}{k_i}.$$

In the directed network, the in/out selectivity of the node *i* is defined as:

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}.$$

Research included aligned parallel Serbian-English collection of documents. Test collection consisted of 50 documents, which are supplied by metadata and keywords. Keywords are annotated by human experts, the authors of articles. It is extracted from Biblisha Digital library http://jerteh.rs/biblisha/ListaDokumenata.aspx?JCID=2&lng=sr [8]. The parallel corpus was cleaned and prepared as a new data resource for research.

Our approach consisted in creating a network from the text and ranking keyword candidates with the highest in/out selectivity values. At first, we did research with the original texts, and then we used texts which we previously stemmed (for English) and lemmatized (for Serbian) by using morphological dictionary for Serbian language [5, 6]. The research showed much better results when we used NLP tools. We also extracted simple words as keyword candidates, and furthermore we expand extraction to word-tuples (two and three words) ranked with the highest selectivity values.

We observed that there was a certain number of words in the original keyword, which do not appear in the texts, and because of that we did evaluation so that we excluded the words which were out of vocabulary. This act provided improvement in results of evaluation, but not to the extent as improvement which came from the use of lemmatization.

In our work, we used python and its module NetworkX, and C# also.

3. Description of the main results obtained

Main results could be outlined as: improvement of existing application for KW extraction and their visualisation, bilingual text collection preparation, application on improved application on new text collection and extraction of bilingual keywords (in Serbian and English). The conducted evaluation on our cases show promising results for further application and integration in other software solutions. The outline of the research paper with key results of this STSM is drafted with the plan to continue cooperation of our teams, prepare and submit joined paper for the third KEYSTONE conference (IKC 2017).

4. Future collaboration with the host institution (if applicable)

We plan to continue our collaboration on this research. We plan to submit the co-authored paper to the IKC 2017 conference. Next, we plan to extent the proposed approach and test the method using longer texts, exactly the full articles in Serbian and English language, which are supplied by keywords given by the author also. We will analyse the results obtained on long texts, and explore how the length of texts effects on the keyword extraction via SBKE method. We will pursuit further collaboration ina form of joint publication in a journal.

5. Foreseen publications/articles resulting from the STSM (if applicable)

(i) Paper for the second 2017 International KEYSTONE conference (IKC 2017).

(ii) We plan to extend the aforementioned paper with further experimental results with the full articles and domain-specific techniques; the extended version will be submitted to a suitable journal, and possibly (if suitable) to a relevant Semantic Web conference.

6. Other comments (if any)

The STSM has provided an opportunity of close interaction, research work, exchange of experience and learn about research on ontologies to represent lexical knowledge, based on methodology developed in previous STSMs by Ana Meštrović. This knowledge, together with new methods for extracting words and phrases from the texts will be helpful for work on my PhD thesis.

Selected references

[1]. S. Beliga, A. Meštrović, S. Martinčić-Ipšić. "Selectivity-Based Keyword Extraction Method". International Journal on Semantic Web and Information Systems (IJSWIS), vol. 12, No. 3, pp. 1-26, 2016, doi: 10.4018/IJSWIS.2016070101

[2]. S. Beliga, A. Meštrović, S. Martinčić-Ipšić. "An Overview of Graph-Based Keyword Extraction Methods and Approaches". Journal of Information and Organizational Sciences, vol. 39, No 1, pages 1-20, 2015.

[3]. S. Beliga, A. Meštrović, S. Martinčić-Ipšić. "Toward Selectivity-Based Keyword Extraction for Croatian News". CEUR Proceedings of the Workshop on Surfacing the Deep and the Social Web (SDSW 2014), Vol. 1310, pp. 1-8, Riva del Garda, Trentino, Italy, 2014.

[4]. S. Beliga, S. Martinčić-Ipšić. "Network-Enabled Keyword Extraction for Under-Resourced Languages". IKC 2016, Cluj-Napoca, Romania, 2016.

[5]. Cvetana Krstev, Ranka Stanković, Ivan Obradović, Biljana Lazić “Terminology Acquisition and Description Using Lexical Resources and Local Grammars”, in Proceedings of the 11th Conference on Terminology and Artificial Intelligence, Granada, Spain, 2015, eds. Thierry Poibeau and Pamela Faber, LexiCon (Universidad de Granada), pp. 81-89, CEUR Workshop Proceedings, ISSN 1613-0073, urn:nbn:de:0074-1495-6,

[6]. Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac, “Rule-based Automatic Multi-word Term Extraction and Lemmatization”, Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23--28 May 2016, 2016, eds. Nicoletta Calzolari et al., ISBN 978-2-9517408-9-1. Zbornik/Proceedings

[7]. Ranka Stanković, Cvetana Krstev, Ivan Obradović, Olivera Kitanović, “Indexing of Textual Databases Based on Lexical Resources: - A Case Study for Serbian”, in Semantic Keyword-Based Search on Structured Data Sources - First COST Action IC1302 International KEYSTONE Conference, IKC 2015, Coimbra, Portugal, September 8-9, 2015. Revised Selected Papers, Springer, LNCS 9398, ISBN 978-3-319-27932-9, DOI 10.1007/978-3-319-27932-9_15, pp. 167-181, 2015

[8]. Ranka Stankovic, Cvetana Krstev, Dusko Vitas, Nikola Vulovic and Olivera Kitanovic. Keyword-based search on bilingual digital libraries, Second COST Action IC1302 International KEYSTONE Conference, IKC 2016, Cluj-de-Napoca, Romania, 8-9 September 2016.

[9]. A. Masucci and G. Rodgers. Differences between normal and shuffled texts: structural properties of weighted networks. Advances in Complex Systems, 12(01):113129 (2009)