Keystone IC1302

STSM Report

STSM Reference: COST-STSMIC1302-36988

Topic: Natural Language Processing Keyword Search for Related Languages

Host: Dr. Ranka Stankovic, University of Belgrade, Faculty of Mining and Geology, Belgrade (RS), ranka.stankovic@rgf.bg.ac.rs

Beneficiary: Dr. Velislava Stoykova, Institute for Bulgarian Language, Sofia (BG), vstoykova@yahoo.com

Duration: 2017-03-26 to 2017-04-02

Purpose:

The purpose of our proposal for STSM allocated to University of Belgrade is to perform a collaboration within the network activities of COST IC1302 Action „semantic KEYword-based Search on sTructured data sOurcEs" – KEYSTONE. The main topic of the research visit "Natural Language Processing Keyword Search for Related Languages" is in the scope of the Action and is aimed to summarize the research experience of both University of Belgrade and Bulgarian Academy of Science, Institute for Bulgarian Language in searching massive amount of lexical data to extract types of semantic relations by using keywords.

Our preliminary research results show that related languages (like Serbian and Bulgarian) share similar lexical and grammatical features and have been presented formally using similar formalisms for lexical knowledge representation like WordNet.

Hence, we expect that they can share comparable results for parallel bilingual keyword search which can give some useful results for Machine Translation in extraction similar semantic relations.

We plan to apply diverse types of statistical search over structured lexical data for both languages and to account how they affect received results. We plan, also, to classify our research results with respect to types of extracted semantic relations and to formulate our conclusions. The main goal is to contribute to the KEYSTONE's conclusive results by publishing related common work.

We hope that the topic of proposed STSM and related collaboration between University of Belgrade and Bulgarian Academy of Sciences will foster the KEYSTONE network activities and will contribute toward better final results of the Action.

Description of the work:

During the time of STSM several tasks were completed at the Department of Mining and Geology, University of Belgrade by Dr. Stoykova, Dr. Stanković and her staff from the department (Biliana Lazić and others):

A. The evaluation of approach and related keyword search methodology including:

(i) Decisions about types of lexical data which are going to be used and about keyword search methodology. We have decided to use multilingual comparable electronic text corpora and the Sketch Engine statistical software for related keyword search experiments. That methodology allows comparison of keyword search results for several languages.

(ii) Decisions about types of semantic relations which we are going to be used – generally synonymic relations of types: and/or, part_of, etc., which present basic semantic relations and allow comparison of keyword search results for several languages.

(iii) Decisions about types of keywords search queries (single word, compound words, language expressions with prepositions or personal names).

(iv) Decisions about statistical search methodology: generation of keywords (by using keyness score), generation of concordances, generation of collocations, and comparison of collocations.

B. For testing our hypotheses, we have created three comparable corpora consisting of mathematical texts in Bulgarian, Serbian and Croatian language by using pipeline methodology of crowd sourcing, at about 100 000 words, respectively.

(i) The three corpora were tagged for synonymic semantic relations (and/or, part_of, etc.)

(ii) The three corpora were processed to search for: keywords (by using keyness score), concordances, collocations, and comparison of collocations.

Results Obtained:

The keyword search results obtained from related three comparable multilingual corpora were similar.

(i) The query search for keywords (by using keyness score) gives comparable results for Bulgarian, Serbian and Croatian language: *mathematics* (BG – математика; SR – matematička; HR - matematika), *geometry* (BG – геометрия, SR - geometrije, HR - geometrije), *vectors*, etc.

(ii) The query search for collocations candidates of related keyword *mathematics* (*pure*, *discrete*, etc.) also gives similar results for Bulgarian, Serbian and Croatian language (BG – чиста, дискретна; SR – čista, diskretna; HR – numeričke)

(iii) The comparison of collocations (which give synonymic relations between two different keywords with common collocations candidates keyword) also gives comparable results for Bulgarian, Serbian and Croatian language.

(iv) The comparison of related multilingual word sketches of a related keyword for Bulgarian, Serbian and Croatian language are also similar, and additionally improve results obtained by comparing collocations.

The similar multilingual keywords search results obtained by accepted methodology confirms that related languages share similar lexical semantic features. The methodology for keyword search of comparable corpora for related languages can give some useful results for Machine Translation in extraction similar semantic relations, instead of using parallel corpora or in lack of them.

The results, also, were presented and discussed at the meeting of JePTex (The Society for Language Resources and Technology of Serbia) on 31.03.2017 as a lecture at the Faculty of Mathematics, University of Belgrade.

http://jerteh.rs/?p=1095

Future Work (Research):

It would be interesting to test the approach used for more languages including non-related ones. We, also, plan to enlarge that methodology for knowledge extraction, since it offers semi-automatic generation of thesauri.

Foreseen Publications:

We plan to summarize obtained results and to offer them to be published and be available for a wider academic audience by submitting a paper to IKC 2017.