STSM Scientific Report

Aleksandar Petrovski
Filozofski fakultet Sveuciliste u Zagrebu, Zagreb, Croatia, 30 May – 17 June 2016

1. Purpose of the STSM

The purpose of this STSM was to extract Multi-Word Expressions (MWEs) from parallel Macedonian and Croatian corpora.

2. Work carried out

The MWEs were extracted from two sources of data:

- Parallel Macedonian / Croatian corpora, sentence aligned, consisting of 500,000+ words
- Wikipedia article titles

2.1 Extracting MWEs from parallel corpora

The corpora were first converted from their original tmx format into two Unicode txt files (one per language), containing 39,725 rows (sentences) each. The textual files were pre-processed: all punctuation marks, symbols and tabs were removed.

In order to extract MWEs from the corpora, two tools have been used:

- Pialign[1]
- Anymalign[2]

They allow creating phrase tables and word alignments, using the pre-processed files as an input. After several unexpected adjustments regarding the input and output formats (particularly utf8 encoding of Cyrillic characters), the results obtained were 313,412 and 1,126,874 aligned phrases, respectively. These phrase tables can be used directly to train the translation models of statistical machine translation tools, like Moses, but that was not our goal. Since both tools are purely statistical, the obtained sequences are not chunks (i.e. don't necessarily exhibit any syntactic or semantic cohesion), but are just ordinary sequences of words which occur in the texts. The vast majority of these sequences are not MWEs, but some of them are. For example, the next row from the output file is an idiom:

На градите ми налегнала санта мраз ||| Na prsa mi naliježe santa leda ||| 1 1 9.20277e-08 2.44854e-31 7.7125e-07 3.33977e-07 2.71828

На градите ми налегнала санта мраз / Na prsa mi naliježe santa leda / *On chest me overlaid drift ice / On my chest overlaid iceberg / I have difficult time, something is pressing me* is the chunk.

In this particular case the composition of the idiom and its meaning are completely identical for Macedonian and Croatian. The numerical values are information about the certainty of

[1] http://www.phontron.com/pialign/
[2] https://anymalign.limsi.fr/

co-occurrence of elements of section and can be used e.g. as the additional information in training the Moses' translation model.

In our case, with Pialign phrases were much better aligned and the next steps were taken with its output sequences only. All rows, not containing at least one space, were removed. The remaining 241,313 phrases might contain MWEs. They will be further filtered using NooJ syntactic grammars.

2.2 Extracting MWEs from Wikipedia titles

Wikipedia publishes database dumps of their content periodically. Two files from the dumps were used:

- the base per-page data, which includes article IDs and their titles in Macedonian
- the inter-language link records

After running a script which extracts a table with parallel Macedonian and Croatian titles, a table with 35,338 rows was obtained. Next, the table was cleansed using regular expressions and manual filtering, which decreased the number of rows to 13,620. This table was sent to the students for further manual cleansing and classification.

Most of the MWEs obtained from Wikipedia article titles are named entities, but there are also many collocations and terms.

An example from the table of a term:

Црн дроб ||| Jetra

In Macedonian, *Црн дроб / Black intestines / Liver* is a MWE, its Croatian translation *Jetra* is a single word.

Another example where a MWE in Croatian is a single word in Macedonian:

Аеродром (Airport) ||| Zračna luka (Air port)

Most of the MWEs contain the same number of single words in both languages:

Алпско скијање (Alpine skiing) ||| Alpsko skijanje (Alpine skiing)

An example where a MWE consists of 3 single words in Macedonian and 2 in Croatian:

Апарати за домаќинството (Appliances for household) ||| Kućanski aparati (Household appliances)

3. Main results obtained

Two parallel Macedonian / Croatian lexicons of MWEs have been obtained. They need additional filtering, but are good starting point for building a parallel lexicon which will contain several thousand entries.

The extracted phrase table by Pialign might be used as an input for training Macedonian / Croatian translation model in Moses.

The method tested from Wikipedia as a source can be treated as language independent in the part before language specific filters are used, and in this respect it could be tested in other

language pairs. Moreover, these lists can be processed with the existing NERC tools (if they exist for a given languages), and be further reduced to generic MWEs. It is sufficient that the NERC tool exists only for one language since the translation equivalence encoded in Wikipedia multilingual links is preserved. In this way a vast multilingual network of MWEs can be built up using by the language technologies the most supported language as a pivotal or hub language.

4. Future collaboration with host institution

The collaboration with the host institution will continue. We are expecting its students to filter and classify potential MWEs from Wikipedia. Both sides will filter the phrase tables extracted by Pialign using NooJ syntactical grammars, each one for her/his own language. The building of multilingual MWE network with NEs filtered out will continue involving other languages.

5. Foreseen publications to result from the STSM

A couple of papers is foreseen as the results of research withing this STSM and they will be sent either to the prominent journals (e.g. Journal of Language Resources and Evaluation, etc.) or conferences/workshops in the field (e.g. TSD, RANLP, LTC, (E)ACL, etc.).

The lexicon of MWEs can be published as the language resource at the META-SHARE platform and in this way COST-PARSEME action and one of its outcomes will be more visible and accessible. The hosting institution runs a node in META-SHARE and can store this language resource in its repository.

6. Confirmation by the host institution of the successful execution of the STSM

Judging by the presented results, the STSM by Aleksandar Petrovski can be considered as very successful for both parties: the invited and the inviting one. The extraction method tested on Wikipedia dumps can be applied to other language pairs. In this way the outcome of this STSM, directly connects to the common work in PARSEME Working Group 3, where methods for MWE resources and their building/filling are being developed.