Report: Short Term Scientific Mission

COST-STSM-ECOST-STSM-IC1207-020314-042515

Nikola Ljubešić

May 14, 2014

1 Purpose of the STSM

The main purpose of the two and a half months STSM was to automatically construct recall-oriented MWE resources for Croatian, Serbian and Slovene, three South Slavic languages with a very limited availability of MWE resources.

Until now, the only freely available MWE resource for Croatian was the Automatic collocation dictionary (ACD_hr)[1] with 32,218 entries and 182,825 collocations. For Slovene, the only freely available resource containing MWEs was the manually edited *Leksikalna baza za slovenščino* (LBS)[2] with 2,500 entries and 45,000 collocations. To best of our knowledge, there was no freely available MWE resource for Serbian.

Both existing resources were built by extracting candidates conforming to pre-specified morphosyntactic patterns. Due to the fact that Slavic languages have free word order, the primary goal of this STSM was to enhance existing methods (both aforementioned resources were compiled with the proprietary SketchEngine) by extracting MWEs from dependency trees. Apart from extracting MWEs from larger corpora and from a higher and more appropriate linguistic abstraction level, which will produce better resources both in terms of size and quality (if constructed automatically), the goal of the STSM was to produce a non-proprietary tool that will enable building similar MWE resources for other free word order languages from parsed corpora.

During the STSM we primarily focused on statistically idiosyncratic MWEs but also made initial explorations of identifying semantic idiosyncrasy with distributional methods.

[1] http://meta-share.ffzg.hr/repository/browse/croatian-automatic-collocations-dictionary/3522fe4a703d11e28a985ef2e4e6c59e70fd1df289e94088b13ac5f01449b1cb/

[2] http://www.slovenscina.eu/spletni-slovar/leksikalna-baza

2 Description of the work accomplished during the STSM

The main tasks carried out during this STSM were the following:

1. preparing / parsing corpora of the languages in question
2. inspecting existing tools for subtree extraction and designing and developing a tool for extracting MWEs based on dependency patterns, either by using existing tools or by constructing a new tool from scratch
3. writing grammars for the languages of interest
4. extracting MWE candidates from the corpora and scoring them statistically
5. considering further possibilities:
    - scoring MWE candidates by semantic transparency using distributional methods
    - extracting examples from corpora for human usage and MWE assessment

2.1 Preparation of the corpora

The main goal was to extract lexicons from large (around 1 billion tokens up) parsed corpora to get high coverage and good statistical idiosyncrasy estimates. It should be noted that parsing such large corpora is a lengthly process, taking upwards of a month on tens of server-grade processor cores with multiple data optimizations like sentence deduplication and long sentence removal.

Most processing power was put into parsing the Croatian hrWaC web corpus containing 1.9 billion tokens. It was successfully parsed during the first half of the STSM.

For Serbian, the srWaC web corpus (890 million tokens) is still being parsed and the final Serbian lexicon will be extracted from that corpus once the parsing process is finished.

For Slovene, the KRES corpus (130 million tokens) was already parsed by the Slovene colleagues and was therefore used in the process of developing the extraction tool. The final Slovene MWE lexicon will be extracted from the GigaFida corpus (1.2 billion tokens) which is still being parsed.

2.2 Development of the tool for MWE extraction

There is a rather small number of tools available for extracting syntactic subtrees from trees (Tigersearch, Tgrep2, Tregex, BTRED/NTRED) and

most of them operate on phrase structure trees. The only tool that can process dependency trees is BTRED/NTRED, but it is strongly tailored for processing the PDT. This is why we decided to develop the necessary tool from scratch. During the first third of the STSM the DepMWEx [depmueks] tool was developed.

The basic idea behind the tool was to enable the user to define dependency patterns as search trees – hierarchical structures whose nodes are boolean functions. Over each parse tree from the corpus an exhaustive search with each search tree is carried out by DepMWEx.

An example search tree is constructed with the following code:

```
V_DO=Tree()
V_DO.add_node(0,V_DO,lambda x:x.data['deprel']=='Pred'))
V_DO.add_node(1,0,lambda x:x.data['deprel']=='Obj' and re.match
    (r'Nc..a',x.data['msd'])!=None)
```

The given search tree consists of two nodes: The root node contains the boolean function checking for the candidate node to be the predicate of the sentence. The child node contains the boolean function defining that the child has to be a direct object of the predicate.

The tool is written in Python and at this point no effort was put into defining a more linguist-friendly syntax for defining the boolean functions. We will tackle this during the development of the next version of the tool once the expressive needs of the search trees will be better known.

The developed formalism is quite expressive. Search trees can (and do), for example, specify that the node agrees in gender, number and case with its parent. As another example, the criteria can include one node's children (which is used to disregard genitive noun phrases having numbers as dependents in Croatian and Serbian).

During the development of the tool, extensive discussions were carried out with Simon Krek (Jožef Stefan Institute), Kaja Dobrovoljc (Trojina Institute) and Tomaž Erjavec. Directions for further development of the tool were discussed as well. Two main ideas emerged:

1. combining a smaller number of dependency patterns in a recursive manner – simple patterns like `(verb (direct object))` and `(verb (prepositional phrase))` could be merged into more complex patterns like `(verb (direct object) (prepositional phrase))` and

2. memorizing the distribution of the dependents (on any abstraction level) of each search tree leaf node enabling later extensions of the extracted candidate (Slovene `(predicate (preposition (noun)))` *dobiti za dan* (eng. *get for day*) has almost exclusively one additional dependent on the noun, the adjective *rojsten* (eng. *get for birthday*)).

2.3 Grammar development

The grammar development and validation was performed during the second third of the STSM.

In collaboration with Kaja Dobrovoljc the existing morphosyntactic patterns for Slovene were rewritten for the newly developed formalism. On top of that, additional patterns were added whose extracts were of very low quality if applied in the morphosyntax-level formalism (like the `(verb (direct object) (prepositional phrase))` or the `(verb (subject) (direct object))` patterns). The final Slovene grammar consists of 75 dependency patterns.

By exploiting the similarity of the three languages, although the dependency parsing formalisms used on Slovene on one side and Croatian and Serbian on the other are rather different, the Slovene grammar was adapted for Croatian and Serbian yielding 62 dependency patterns.

During the pattern adaption and writing process, the pattern candidates were validated on the KRES corpus and a 10% portion of the hrWaC corpus.

2.4 Extraction of MWEs

The last third of the STSM was dedicated to extracting and scoring the MWE candidates from the prepared corpora. For all three languages, each of the extracted MWE candidates was scored by its frequency (with a minimum frequency filter of 5) and the logDice association score.

For Croatian, a lexicon from all the 1.9 billion tokens of the hrWaC corpus was extracted. To ensure the quality of the resulting lexicon, a lemma filter consisting of all non-proper-nouns from the Croatian morphological lexicon (HML) and the Apertium lexicon was used. Namely, the lemmatization procedure used on the hrWaC and srWaC corpora is statistical and trained over 130k tokens of running text only.

Evaluation of the extracts from hrWaC is under way in collaboration with Darja Fišer (Faculty of Arts, University of Ljubljana) and her Croatian doctoral student Marina Peršurić. By annotating the top 20 (sorted by logDice) MWE candidates of each dependency pattern for 12 sample lexemes, we will estimate the accuracy of each dependency pattern given the frequency of the headword and the frequency and logDice values of the MWE candidate. These estimates will be used in refining the current selection of MWE candidates included in the resource. If the annotation efforts result in a larger annotated sample, learning a supervised filtering model from the annotated data will be taken into consideration.

For Serbian, a provisional lexicon was extracted from 40% of the parsed srWaC corpus. No lemma filter was applied due to the lack of such resources. For the final lexicon that will be built from the fully parsed srWaC corpus we hope to apply a supervised approach of classifying lemmas as "good"

or "bad" by using features like the distribution of morphosyntactic tags, lemma endings etc.

For Slovene, a provisional lexicon was extracted from the KRES corpus. A final lexicon will be extracted from the GigaFida corpus once it is parsed. There is no need to apply a lemma filter to the Slovene extract because the preprocessing of Slovene is of higher accuracy than for the two other languages.

2.5 Inspecting further possibilities

Two additional tasks were tackled in the last weeks of the STSM: 1. calculating semantic transparency by using distributional methods and 2. extracting good MWE examples from corpora to augment the use of the lexicons by human users, in particular lexicographers.

2.5.1 Semantic transparency via distributional methods

Two initial computations of semantic transparency have been performed. The first has shown no, and the second medium correlation to the phenomenon of interest. We are currently performing new experiments and will continue this line of research after the STSM.

2.5.2 Extracting good examples from corpora

All together 11 feature extractors for example quality assessment were written (length of the sentence, syntactic complexity as normalized number of the length of syntactic arcs, normalized number of pronouns in a sentence, normalized number of punctuations etc.). At this point we defined a heuristic that scores each sentence by using the extracted features. Our plan is to encode 10 best examples for each of the MWE candidates in the lexicon.

We will continue working on this problem by using the defined features in a supervised learning setting. This line of work will be continued in collaboration with Simon Krek and his colleague Iztok Kosem (Trojina Institute).

3 Description of main results obtained

The main results of this STSM are the three lexical resources and the tool used for building those resources.

The Croatian MWE lexicon hrMWELex version 0.5 (extracted from the hrWaC corpus) is published on

`http://nlp.ffzg.hr/resources/lexicons/hrmwelex/`

under the CC-BY-SA 3.0 license. Further versions of the lexicon updated with corpus examples and semantic transparency information will be published on the same location.

The Serbian MWE lexicon srMWELex version 0.5 (extracted from 40% of the srWaC corpus) is published on

`http://nlp.ffzg.hr/resources/lexicons/srmwelex/`

under the CC-BY-SA 3.0 license. Once the parsing of the srWaC corpus has finished, the updated resource will be available from the same location.

The Slovene lexicon slMWELex version 0.5 (at this point extracted from KRES) is published on

`http://nlp.ffzg.hr/resources/lexicons/slmwelex/`

under the CC-BY-SA 3.0 license. Once the parsing of the Gigafida corpus has finished, the updated resource will be available from the same location.

The size of the currently published lexicons is this:

lexicon	# of headwords	# of MWE candidates
hrMWELex v0.5	46,293	12,750,029
srMWELex v0.5	23,594	3,279,864
slMWELex v0.5	47,579	6,383,963

The first (random) entry from each of the three lexicons is given in the Appendix.

The developed DepMWEx tool, used for building the three lexical resources, is published on

`https://github.com/nljubesi/depmwex`

under the GNU GPL license. It is still in an early prototype stage, but nevertheless fully usable. Along with the tool, the grammars for Croatian (which is applicable to Serbian) and Slovene are published as well.

4 Future collaboration with host institution

The collaboration with the host institution will continue on multiple levels.

First of all, the development of the slMWELex resource will continue as the parsing of the GigaFida corpus is finished.

Secondly, the task of extracting good examples from corpora will be dealt with together as part of the ENeL (IS1305) COST action.

Finally, outside the scope of this action, but largely due to the STSM, collaboration on two projects, one European, one national, has been agreed on.

5 Foreseen publications/articles resulting or to result from the STSM

The hrMWELex lexicon will be described in a paper submitted to the Slovene NLP conference IS-JT2014 taking place in October 2014.

The DepMWE tool is expected to be described in a paper by the end of the year and presented at one of the NLP conferences in 2015.

The two remaining resources, slMWELex and srMWELex will be presented at NLP conferences next year.

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

Tomaž Erjavec, Dept. for Knowledge Technologies, Jožef Stefan Institute: The STSM achieved all its stated goals and the cooperation with dr. Ljubešić was highly beneficial also for JSI and for the advancement of MWE methods and resources for Slovene. The STSM also gave the opportunity for dr. Ljubešić to actively participate in the work of researchers from other Slovene institutions with highly promising results. Furthermore, during his visit he also participated in the preparation of a Slovene basic research project proposal and, if the propsal is sucesfull, will be part of the project team. Given his outstanding work during the STMS, he was also invited to participate in a current EU project in which our department is a partner.

Appendix

First (random) entry from each of the three lexicons, hrMWELex, srMWELex and slMWELex.

```
<entry>
        <lexeme pos="Nc">slamarica</lexeme>
        <pattern type="gbz_u_SBZ4">
                <dependents freq="7" logdice="2.3616">
                        <lexeme pos="Vm">sakriti</lexeme>
                </dependents>
        </pattern>
        <pattern type="gbz_na_SBZ5">
                <dependents freq="15" logdice="2.56527">
                        <lexeme pos="Vm">spavati</lexeme>
                </dependents>
        </pattern>
</entry>
```

```
<entry>
        <lexeme pos="Ap">podmlađen</lexeme>
        <pattern type="PBZ0_sbz0">
                <dependents freq="10" logdice="3.03931">
                        <lexeme pos="Nc">sastav</lexeme>
                </dependents>
                <dependents freq="5" logdice="1.51357">
                        <lexeme pos="Nc">ekipa</lexeme>
                </dependents>
                <dependents freq="8" logdice="1.01297">
                        <lexeme pos="Nc">tim</lexeme>
                </dependents>
        </pattern>
        <pattern type="rbz_PBZ0">
                <dependents freq="13" logdice="4.42708">
                        <lexeme pos="Rg">znatno</lexeme>
                </dependents>
                <dependents freq="9" logdice="3.21855">
                        <lexeme pos="Rg">dosta</lexeme>
                </dependents>
        </pattern>
</entry>

<entry>
        <lexeme pos="So">glavonožec</lexeme>
        <pattern type="zveze_s_predlogi">
                <dependents freq="2" logdice="2.488">
                        <lexeme pos="Dt">med</lexeme>
                </dependents>
                <dependents freq="3" logdice="-0.65248">
                        <lexeme pos="Do">med</lexeme>
                </dependents>
                <dependents freq="3" logdice="-1.43965">
                        <lexeme pos="Dm">pri</lexeme>
                </dependents>
                <dependents freq="2" logdice="-3.98979">
                        <lexeme pos="Dt">za</lexeme>
                </dependents>
        </pattern>
        <pattern type="SBZ0_in/ali_SBZ0">
                <dependents freq="3" logdice="5.09511">
                        <lexeme pos="So">školjka</lexeme>
                </dependents>
```

```
                <dependents freq="14" logdice="5.05351">
                        <lexeme pos="So">riba</lexeme>
                </dependents>
                <dependents freq="2" logdice="2.41105">
                        <lexeme pos="So">rak</lexeme>
                </dependents>
        </pattern>
        <pattern type="sbz0_SBZ2">
                <dependents freq="3" logdice="6.55154">
                        <lexeme pos="So">pojavljanje</lexeme>
                </dependents>
                <dependents freq="4" logdice="0.49343">
                        <lexeme pos="So">vrsta</lexeme>
                </dependents>
        </pattern>
        <pattern type="SBZ1_gbz">
                <dependents freq="2" logdice="0.62867">
                        <lexeme pos="Gg">stati</lexeme>
                </dependents>
                <dependents freq="3" logdice="-5.1403">
                        <lexeme pos="Gp">biti</lexeme>
                </dependents>
        </pattern>
        <pattern type="gbz_SBZ4">
                <dependents freq="3" logdice="5.13066">
                        <lexeme pos="Gg">risati</lexeme>
                </dependents>
        </pattern>
        <pattern type="gbz_pri_SBZ5">
                <dependents freq="2" logdice="1.87952">
                        <lexeme pos="Gg">razviti</lexeme>
                </dependents>
        </pattern>
</entry>
```