"Kad se mnogo malih složi": Collaborative development of gold resources for Slovene, Croatian and Serbian

Nikola Ljubešić[1], Tanja Samardžić[2], Tomaž Erjavec[1], Darja Fišer[3,1], Maja Miličević[4], Simon Krek[5]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2] URPP Language and Space, University of Zurich
[3]Department of Translation, University of Ljubljana
[4] Deptartment of Linguistics, Faculty of Philology, University of Belgrade
[5] Centre for Language Resources and Technologies, University of Ljubljana

SlaviCorp 2018, Prague, Czech Republic
September 24, 2018

Introduction

Overview

- Textual datasets manually annotated with linguistic information are a prerequisite for developing language technologies based on machine learning
- The talk presents a series of collaborations between researchers developing such datasets for Slovene, Croatian and Serbian
- Close relatedness of these languages: opportunity for a synchronized approach
- Complex political situation did not allow top-down development: so, an informal, bottom-up approach
- Main synergistic effect:
  drastically lowering the effort to produce datasets:
  1. Development of annotation guidelines
  2. Setting up the technical infrastructure for annotation
  3. Pre-annotation of data with models trained for another, but very close language

Kad se mnogo malih složi, tad se snaga stoput množi

- The main song from a legendary children's movie from 1976 "Vlak u snijegu"
- A train full of children gets stuck in snow, with joint efforts they enable the train to continue and bring them back to their village

Morphosyntax

Morphosyntactic specifications

MULTEXT-East

- MULTEXT-East specifications define morphosyntactic features and tagsets for 16 languages: 11 Slavic
- MULTEXT-East resource also contain lexicons and an annotated corpus which follow the specifications

Croatian and Serbian MULTEXT-East V5

- Croatian was the only language where only the specifications were available: without a lexicon and corpus, rather useless
- We produced a new version of Croatian specifications, which borrows heavily from the Slovene approach
- Serbian also updated, to be in line with the new Croatian specifications
- The new specifications (MULTEXT-East Version 5) are still a draft: `http://nl.ijs.si/ME/V5/msd/`

Lexicons and Corpora

Inflectional lexicons

- Compilation of lexicons financed by Abu-MaTran (FP7)
- Produced infl. lexicons for Croatian (hrLex) and Serbian (srLex)
- They were developed via machine learning ranking of lemma and paradigm candidates for out-of-vocabulary words
- Heavy re-use of paradigms and lexemes

Tagger training data

- Manual corpus annotation financed by the ReLDI (Swiss Science Foundation) project
- Annotated SETimes.SR (87k tokens)
- Pre-annotated with a model trained on the parallel SETimes.HR Croatian dataset, then manually corrected

Syntax

Syntax

Universal Dependencies

- Financed by ReLDI
- Universal Dependency annotation added to the SETimes.SR dataset
- Pre-annotation with a model trained on the parallel SETimes.HR Croatian dataset
- Only 15% of tokens required manual correction

Coordination of Slovene, Croatian and Serbian UDs

The UD annotation effort for Croatian and Serbian on one side and Slovene on the other are currently not coordinated, planned as an additional synergy in the future

Social media

Processing of social media texts

Janes project

Slovene national project, manually annotated datasets for processing Slovene user generated content (UGC): tokenisation, sentence splitting, normalization, morphosyntactic tagging, lemmatisation and named entities.

ReLDI project: adding Croatian and Serbian

Annotation guidelines translated into Serbian and additional campaigns in WebAnno for Croatian and Serbian UGC.

High complexity of the annotation campaign: reusing the annotation guidelines and the annotation technology drastically lowered the costs of producing these datasets.

Named entities

Named entity recognition

Slovene

- Slovene dataset ssj500k was partially annotated with NEs, but no annotation guidelines
- We wrote the guidelines, corrected existing and annotated additional sentences with NEs
- Annotation guidelines extended to UGC and historical language.

Croatian and Serbian

- Supported by ReLDI
- Croatian dataset hr500k annotated by following the same guidelines and using the same annotation technology
- Serbian dataset SETimes.SR pre-annotated with the Croatian model and manually corrected

Semantic roles

Semantic role labeling

Slovene and Croatian datasets

- Bilateral Slovene-Croatian project on collaborative development of semantic role labeling for Croatian and Slovene
- Joint annotation guidelines developed and annotation campaigns run simultaneously on Slovene ssj500k and Croatian hr500k datasets, using the same annotation technology
- Issues in each language were jointly discussed and resolved

Extension to Serbian

As with morphosyntax, syntax and named entities, a future task is to pre-annotate the Serbian SETimes.SR dataset with a model trained on the parallel SETimes.HR Croatian dataset.

Coreference

Coreference resolution

Serbian as starting language

- In most projects transfer went either from Slovene and/or from Croatian. Here, Serbian is the initial language
- Annotation currently applied to the Serbian gold standard SETimes.SR corpus, with detailed annotation guidelines developed

Extension to remaining languages

Our plan for the near future is for the annotation guidelines to be transferred to Croatian and Slovene, where the Croatian data will be pre-annotated with the Serbian model.

Conclusions

Conclusions

- Described an approach of borrowing (1) annotation guidelines, (2) technology for manual annotation and (3) existing models for pre-annotating closely related languages, i.e. Slovene, Croatian and Serbian.
- Started as informal, non-funded endeavor, continued by orchestrating activities inside different (and only partially related) projects
- We plan to continue this work as cooperation between the Slovene research infrastructure CLARIN.SI and the cross-national ReLDI centre for language data (follow-up of the ReLDI project)

"Kad se mnogo malih složi": Collaborative development of gold resources for Slovene, Croatian and Serbian

Nikola Ljubešić[1], Tanja Samardžić[2], Tomaž Erjavec[1], Darja Fišer[3,1], Maja Miličević[4], Simon Krek[5]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2] URPP Language and Space, University of Zurich
[3]Department of Translation, University of Ljubljana
[4] Deptartment of Linguistics, Faculty of Philology, University of Belgrade
[5] Centre for Language Resources and Technologies, University of Ljubljana

SlaviCorp 2018, Prague, Czech Republic
September 24, 2018