# Slovene-Croatian Treebank Transfer Using Bilingual Lexicon Improves Croatian Dependency Parsing

Željko Agić, Danijela Merkler, Daša Berović

Faculty of Humanities and Social Sciences
University of Zagreb

IS-JT 2012, Ljubljana, 2012-10-08

# Motivation

- dependency treebanks of Croatian and Slovene are relatively small, but sufficient to perform parsing experiments
- transferring a treebank from Slovene to Croatian in order to improve Croatian parsers
- treebank transfer
  - translating a treebank from source language to target language while maintaining its syntactic annotation layer
  - source language: Slovene, target language: Croatian
- relatedness of Croatian and Slovene
  - syntactic transfer method based on a Croatian-Slovene bilingual lexicon might improve dependency parsing scores

# Transfer

- requires a dependency treebank of Slovene and a dictionary
- Slovene Dependency Treebank (SDT)
  - a part of the morphosyntactically annotated Slovene 1984 corpus from Multext-East
  - approx. 30,000 tokens in 2,000 sentences
  - JOS corpus not compatible with PDT-style syntactic functions
- bilingual lexicon
  - constructed from the Croatian-Slovene subset from the 1984. parallel corpus
  - sentence-aligned, keeping only 1:1 sentence alignments
  - constructed from 6,337 sentence pairs using GIZA++
  - contains 52,502 Slovene-Croatian word pairs
  - entries were sorted by translation probability obtained from the parallel corpus

# Experiment

- three stages
  - translation of SDT to Croatian (hr-SDT)
  - assigning the Croatian metadata to hr-SDT
  - training and testing parsers
    - manually dependency parsed Croatian texts — Croatian Dependency Treebank (HOBS)
    - merging HOBS and hr-SDT
- translation of SDT to hr-SDT
  - word pairs with highest probability chosen from the dictionary
  - assessing translation quality
    - 100 randomly selected sentences were manually evaluated for adequacy and fluency on 1-5 scale
    - adequacy 3.64, fluency 2.99, BLEU 0.1962
- assigning the Croatian metadata to hr-SDT
  - (not) keeping Slovene MSD-tags, (not) translating lemmas

# Experiment

- training and testing parsers
  - MSTParser
    - state-of-the-art graph-based dependency parser generator
    - used to generate second order arc-factored non-projective parsers for Croatian
    - observed LAS ca 74.53% on HOBS
  - CroDep
    - a novel k-best maximum spanning tree dependency parser with valency lexicon reranking
    - parsing score of approx. 77.21% on HOBS
  - training sets were created by attaching hr-SDT to HOBS training sets
    - 10 disjoint testing sets of approx. 5,000 tokens and 10 disjoint training sets of approx. 83,000 tokens from HOBS
    - each training set was merged with both versions of hr-SDT
    - 1984 test set created manually by annotating 345 sentences from the Croatian 1984 corpus

# Results

| Test set | Model | MST | CroDep |
|---|---|---|---|
| hr-1984 | HOBS | 68.51 | 71.37 |
| | HOBS + hr-SDT | 69.44 | 72.26 |
| | HOBS + hr-SDT tagged | 69.69 | 72.48 |
| HOBS | HOBS | 74.53 | 77.21 |
| | HOBS + hr-SDT | 73.96 | 76.77 |
| | HOBS + hr-SDT tagged | 74.00 | 76.89 |

Table: Overall parsing accuracy (LAS)

- usefulness of treebank transfer is domain-dependent
  - introducing hr-SDT, corpus of fictional texts
    - decreases the overall parsing accuracy on newspaper texts
    - improves parsing hr-1984 test set
- in average, CroDep is topping MSTParser by approx. 2.71% LAS across domains

# Conclusions and future work

- treebank transfer between similar languages using bilingual lexicon improves dependency parsing accuracy
  - improvement is domain-dependent
- future work directions
  - domain-specific bilingual lexica
  - translations of higher quality
    - using probabilistic word-by-word decoding
    - construction bilingual lexica using English as interlingua
  - repeating experiment by setting Croatian as source and Slovene as target language
    - mapping syntactic annotations of the JOS corpus to SDT style and vice versa, as well as HOBS
  - other language pairs with compatible treebanks
    - e.g. Czech-Slovene and Czech-Croatian
  - linguistic rules for syntactic transfer in m:n word alignments

Thank you for your attention.