Universal Dependencies for Croatian (that Work for Serbian, too)

Željko Agić*  Nikola Ljubešić†

* Center for Language Technology
  University of Copenhagen, Denmark

† Dept. of Information and Communication Sciences,
  Faculty of Humanities and Social Sciences,
  University of Zagreb

BSNLP 2015, 10th Sep 2015
for parsing we need supervision in form of annotated corpora

dependency treebanks costly to develop and follow different annotation schemes across languages

this hinders cross-lingual parsing and enabling LT for under-resourced languages

Universal Dependencies [Nivre et al., 2015] address this issue by providing homogenous dependency treebanks

parts of speech, morphological features and syntactic annotations across 18 languages

[McDonald et al., 2013] stress the two obvious gains from uniform schemata:

1 more exact evaluation of dependency parsers

2 typologically motivated transfer of dependency parsers to under-resourced languages
Contributions

- focus on cross-lingual dependency parsing of two under-resourced South Slavic languages

1. dependency treebank for Croatian
2. cross-domain test sets for Croatian and Serbian
3. set of experiments for parsing the languages within the UD framework
4. cross-lingual parsing experiments, target Croatian and Serbian by source models from 10 treebanks, two types (CoNLL and UD)
5. make our datasets available under free-culture licensing https://github.com/ffnlp/sethr
The treebank

- built on top of the `Setimes.Hr` dependency treebank [Agić and Ljubešić, 2014]
- 3,557 training sentences (newswire)
- 200 dev sentences from same source
- 400 test sentences
  - 200 Croatian, 200 Serbian
  - 200 from same source, 200 from Wikipedia
  - 100 per source and language
- implement the following annotation layers (first two mandatory):
  1. universal POS tags
  2. dependency attachment
  3. universal morphological features
**Morphology**

- **SETIMES.HR** implements (a revision of) the Multext East version 4 morphosyntactic tagset (MTE4) [Erjavec, 2012]
- manually convert it to
  - UD’s universal POS tags (UPOS)
  - universal morphological features
- out of 17 UPOS tags 14 used in our treebank
- leave out determiners (DET), interjections (INTJ), and symbols (SYM)
- MTE4 abbreviations mapped context-dependent to appropriate UPOS tags, mostly nouns, but adverbs as well ("npr." = "e.g.")
- conflate the 1316 seen tags to 14
- manual annotation by four expert annotators
- apply 39 out of 40 universal relations (leave out the speech-specific *reparandum*)
- 15 syntactic tags of *SETIMES.HR* generalisations of the 39 Croatian UD concepts
- non-projective sentences
  - HOBS [Tadić, 2007] 20%
  - *SETIMES.HR* 10.1%
  - UD 7.6%
Experimental setup

- two sets of experiments
  1. Croatian as source – monolingual parsing of Croatian and transfer to Serbian
  2. Croatian and Serbian as target – transfer of delexicalised parsers from 10 well-resourced languages to Croatian and Serbian

- parser – *mate-tools* graph-based parser of [Bohnet, 2010]
- evaluation – LAS and UAS
- features
  - word form (FORM)
  - coarse-grained POS tag (CPOS)
  - morphological features (FEATS)
  - dependencies (HEAD, DEPREL)
- delexicalised parser drops FORM and FEATS
train on the Croatian train set, evaluate on Croatian and Serbian test sets

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Features</th>
<th>Croatian NEWS</th>
<th>Croatian WIKI</th>
<th>Serbian NEWS</th>
<th>Serbian WIKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set.HR</td>
<td>CPOS</td>
<td>82.2</td>
<td>77.1</td>
<td>80.8</td>
<td>79.8</td>
</tr>
<tr>
<td></td>
<td>+ FEATS</td>
<td>84.3</td>
<td>80.7</td>
<td>83.0</td>
<td>82.6</td>
</tr>
<tr>
<td>UD</td>
<td>CPOS</td>
<td>84.8</td>
<td>80.8</td>
<td>82.4</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>+ FEATS</td>
<td>86.9</td>
<td>84.5</td>
<td>86.0</td>
<td>83.7</td>
</tr>
</tbody>
</table>

- morphological features add consistently 2-4 points
- UD outperforms Setimes.Hr for 2-3 points?
flip POS information to observe the impact of the syntactic layer only

for any final conclusions the parser outputs still have to be evaluated extrinsically on downstream tasks!
Croatian and Serbian as targets

- replicate the single-source delexicalised transfer setups of [McDonald et al., 2011, McDonald et al., 2013] – CPOS the only observable feature
- select 10 languages with treebanks in both CoNLL 2006-2007 and UD v1.0
- evaluate CoNLL on Setimes.HR – heterogenous setting
- UD evaluated on UD – homogenous
- evaluate CoNLL on UAS only as CoNLL and Setimes.HR labels do not overlap
- for CoNLL experiments map the UPOS to [Petrov et al., 2012]
# Croatian and Serbian as targets

<table>
<thead>
<tr>
<th>Source</th>
<th>CoNLL hrv</th>
<th>CoNLL srp</th>
<th>UD hrv UAS</th>
<th>UD hrv LAS</th>
<th>UD srp UAS</th>
<th>UD srp LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>49.8</td>
<td>49.2</td>
<td>64.1</td>
<td>50.6</td>
<td>66.6</td>
<td>53.8</td>
</tr>
<tr>
<td>Czech</td>
<td>36.3</td>
<td>36.1</td>
<td>69.9</td>
<td>54.8</td>
<td>71.9</td>
<td>57.3</td>
</tr>
<tr>
<td>Danish</td>
<td>42.1</td>
<td>42.2</td>
<td>56.7</td>
<td>44.2</td>
<td>56.9</td>
<td>45.6</td>
</tr>
<tr>
<td>German</td>
<td>40.6</td>
<td>41.5</td>
<td>58.1</td>
<td>41.8</td>
<td>60.0</td>
<td>45.1</td>
</tr>
<tr>
<td>Greek</td>
<td>61.7</td>
<td>63.4</td>
<td>52.0</td>
<td>32.8</td>
<td>53.8</td>
<td>35.1</td>
</tr>
<tr>
<td>English</td>
<td>46.3</td>
<td>46.5</td>
<td>54.6</td>
<td>41.3</td>
<td>57.1</td>
<td>44.1</td>
</tr>
<tr>
<td>Spanish</td>
<td>30.4</td>
<td>33.5</td>
<td>60.8</td>
<td>43.7</td>
<td>64.1</td>
<td>47.5</td>
</tr>
<tr>
<td>French</td>
<td>40.3</td>
<td>42.7</td>
<td>56.6</td>
<td>41.4</td>
<td>56.3</td>
<td>42.3</td>
</tr>
<tr>
<td>Italian</td>
<td>43.2</td>
<td>45.0</td>
<td>61.3</td>
<td>45.5</td>
<td>62.5</td>
<td>47.6</td>
</tr>
<tr>
<td>Swedish</td>
<td>40.2</td>
<td>41.2</td>
<td>55.9</td>
<td>42.7</td>
<td>56.4</td>
<td>44.4</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td><strong>43.1</strong></td>
<td><strong>44.1</strong></td>
<td><strong>59.0</strong></td>
<td><strong>43.9</strong></td>
<td><strong>60.6</strong></td>
<td><strong>46.3</strong></td>
</tr>
</tbody>
</table>
Conclusion and future work

- presented the Croatian syntactic dependency treebank within the Universal Dependencies framework
- cca. 4,000 sentences with two-domain two-languages test sets
- intrinsic evaluation via monolingual parsing with \(~80\) LAS on both languages
- although the label set is twice the size, UD proven to be easier to parse than \textsc{Setimes.Hr}
- heterogenous vs. homogenous delexicalised cross-lingual parsing – homogenous gives much better results, following typological similarities
- future work
  - writing UD documentation
  - currently do not utilise language-specific features in neither morphology nor syntax
  - downstream evaluation!
Universal Dependencies for Croatian (that Work for Serbian, too)

Željko Agić* Nikola Ljubešić†

* Center for Language Technology
  University of Copenhagen, Denmark

† Dept. of Information and Communication Sciences,
  Faculty of Humanities and Social Sciences,
  University of Zagreb

BSNLP 2015, 10th Sep 2015