Multilingual projection for parsing truly low-resource languages

Željko Agić    Anders Johannsen    Barbara Plank
Héctor Martínez Alonso    Natalie Schluter    Anders Søgaard

zeag@itu.dk

ACL 2016, Berlin, 2016-08-08
Motivation

Cross-lingual dependency parsing: almost solved?
Motivation

State of the art: +82% UAS on average, using an annotation projection-based approach.
Motivation

(For German, Spanish, French, Italian, Portuguese, and Swedish.)
Motivation

Treebanks are only available for the 1%. Cross-lingual learning aims at enabling the remaining 99%.

http://xkcd.com/688/
Motivation

- The 1% is very cosy.
- Limited evaluation spawns bias.

- POS tagger availability
- parallel corpora: coverage, size, quality of fit
- tokenization
- sentence and word alignment
Motivation

Cross-lingual dependency parsing: almost solved a bit broken.
Our approach

Start simple, but fair.

1. Low-resource languages are low-resource.
2. A handful of resource-rich source languages do exist.
3. Annotation projection seems to work.
4. Go for high coverage of the 99%, evaluate where possible.
Our approach

Projection of POS and dependencies from multiple sources (the 1%) to as many targets (the 99%) as possible.
Our approach

1. Tag and parse the source sides of parallel corpora.
2. For each source-target sentence pair, project POS tags and dependencies to the target tokens.
3. Decode the accumulated annotations, i.e., select the best POS and head for each token among the candidates.
4. Train target-language taggers and parsers.
Our approach

What do we need for it to work?
High-coverage parallel corpora.

- Bible: +1,600 languages online
- Watchtower: +300
- UN Declaration of Human Rights: +500
- OpenSubtitles
Tools

- source-side
  - POS tagger
  - arc-factored dependency parser

- no free preprocessing for parallel corpora
  - simplistic punctuation-based tokenization for all languages
  - automatic sentence and word alignment
Evaluation

Generate models for the many, evaluate for the few.

21 sources, 6 + 21 targets (UD 1.2)
100 models, easily extends to +1000
Our approach

How exactly does our projection work?
Projecting POS

Parallel sentence:

English: In the beginning was the Word

Croatian: U početku bijaše Riječ

German: Am Anfang war das Wort

POS tags:

<table>
<thead>
<tr>
<th>HR</th>
<th>EN</th>
<th>DE</th>
<th>Voted</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>ADP</td>
<td>ADP</td>
<td>ADP</td>
<td>0.8667</td>
</tr>
<tr>
<td>početku</td>
<td>NOUN, DET</td>
<td>NOUN</td>
<td>NOUN</td>
<td>0.7448</td>
</tr>
<tr>
<td>bijaše</td>
<td>VERB</td>
<td>VERB</td>
<td>VERB</td>
<td>0.8560</td>
</tr>
<tr>
<td>Riječ</td>
<td>DET, NOUN</td>
<td>DET, NOUN</td>
<td>NOUN</td>
<td>0.6307</td>
</tr>
</tbody>
</table>
Projecting dependencies

G[Vi] and G[Vj] diagrams with alignment and voting processes.
Projecting dependencies

\[ G[V_t] \]

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
ADP, ADP & & & & & \\
2 & & & & & \\
NOUN, NOUN & & & & & \\
4 & & & & & \\
AUX, AUX & & & & & \\
5 & & & & & \\
DET & & & & & \\
6 & & & & & \\
NOUN, NOUN & & & & & \\
\end{array}
\]

\[ (0.3 \times 0.95 \times 0.85) + (0.5 \times 0.75 \times 0.65) = 0.486 \]

\[ 2) \text{ voting} \]

\[ 3) \text{ decoding} \]

\[ \text{beginning} \rightarrow \text{was} \rightarrow \text{word} \rightarrow \text{beginning} \rightarrow \text{the} \rightarrow \text{in} \]

\[ \text{DMST} \]
Our approach

Our models are built from scratch. The parsers depend on the cross-lingual POS taggers.
Experiment

- baselines
  - multi-source delexicalized transfer
  - DCA projection
  - voting multiple single-source delexicalized parsers
- upper bounds
  - single-best delexicalized parser
  - self-training
  - direct supervision
- parameters
  - parallel corpora: Bible vs. Watchtower
  - word alignment: IBM1 vs. IBM2
Results

Our approach vs. the rest:
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>multi-source</td>
<td>45.43</td>
</tr>
<tr>
<td>reparse</td>
<td>47.79</td>
</tr>
<tr>
<td>DCA</td>
<td>47.87</td>
</tr>
<tr>
<td>our approach</td>
<td>53.47</td>
</tr>
<tr>
<td>single-best</td>
<td>48.52</td>
</tr>
</tbody>
</table>
Results

IBM1 vs. IBM2 at their best:
Results

And the moment you’ve all been waiting for:
Results

parsing
53.47 > 49.57

tagging
70.56 > 65.18
Conclusions

Our approach is simple, and it works.

▶ Take-home messages

1. Limited evaluation spawns benchmarking bias.
2. Go for higher coverage, evaluate on a subset if need be.

3. Simple and generic beat complex and finely tuned.
   ▶ IBM1 vs. IBM2
   ▶ our projection vs. DCA

4. The baselines are better than credited for.
Follow-up work: Wednesday at 15:30 (Session 8D)

Joint projection of POS and dependencies from multiple sources!
Thank you for your attention. 😊

Data freely available at: https://bitbucket.org/lowlands/