KL_{cpos}^3

a Language Similarity Measure for Delexicalized Parser Transfer

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ACL, Beijing, 28 July 2015
Outline

- Intro and motivation
- MSTParser and its delexicalization
- Single-source delexicalized parser transfer
  - $KL_{cpos3}$ for source selection
- Multi-source delexicalized parser transfer
  - $KL_{cpos3}$ for source weighting
- Results
Semi-supervised parsing

- fully supervised dependency parsing
  - requires training data (treebank) or a grammar
  - there are ~100 treebanks (manually annotated)
  - there are ~7,000 languages
  - + various domains, language evolution...

- semi-supervised parsing
  - utilize existing resources, avoid new annotations
    - treebanks for other langs (HamleDT: 30 langs)
    - unannotated data (here: POS tagged)
(Lexicalized) MSTParser

Rudolf NOUN likes VERB trains NOUN

#root

22.3 20.7

-5.2 16.4

48.7

32.5

13.8

7.3 -12.4
(Lexicalized) MSTParser
Delexicalized MSTParser

#root

NOUN

VERB

NOUN

\[ \begin{align*}
\text{NOUN} & \rightarrow 7.4 \rightarrow \text{NOUN} \\
\text{VERB} & \rightarrow -4.3 \rightarrow \text{VERB} \\
\text{NOUN} & \rightarrow 16.1 \rightarrow \text{VERB} \\
\text{VERB} & \rightarrow -1.4 \rightarrow \text{VERB} \\
\text{NOUN} & \rightarrow 9.7 \rightarrow \text{NOUN} \\
\end{align*} \]
Single-source delex parser transfer

- (Zeman and Resnik, 2008)
- train a delexicalized parser on a source language treebank (e.g. Czech – PDT)
- apply it to a target language, without a treebank but with a POS tagger (e.g. Slovak)
Utilizing multiple treebanks

- HamleDT: 30 harmonized treebanks
  - (split: 12 development TBs, 18 testing TBs)
- How do we choose the source treebank?
- Can we use more/all source treebanks?
Utilizing multiple treebanks

- HamleDT: 30 harmonized treebanks
  - (split: 12 development TBs, 18 testing TBs)
- How do we choose the source treebank?
- Can we use more/all source treebanks?

- treebank concatenation (McDonald et al., 2011)
  - if you don't know what to do, just concatenate all the data you have
  - concatenate all source treebanks $\rightarrow$ train a parser $\rightarrow$ apply the parser to the target text
Choosing the source treebank

- src should be as similar to tgt as possible
  - WALS (Naseem et al., 2012)
  - POS $n$-gram model (Søgaard and Wulff, 2012)
Choosing the source treebank

- src should be as similar to tgt as possible
  - WALS (Naseem et al., 2012)
  - POS $n$-gram model (Søgaard and Wulff, 2012)
  - $KL_{cpos_3} (tgt, src)$: Kullback-Leibler divergence of POS trigram distributions
KL_{cpos^3}(tgt, src) = \sum_{\forall cpos^3 \in tgt} f_{tgt}(cpos^3) \cdot \log \left( \frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)} \right)

cpos^3 = \langle cpos_{i-1}, cpos_i, cpos_{i+1} \rangle

f(cpos^3) = \frac{\text{count}(cpos^3)}{|corpus|}
# Sample of results (HamleDT)

<table>
<thead>
<tr>
<th>Target lang.</th>
<th>KL&lt;sub&gt;cpos3&lt;/sub&gt; selected src lang.</th>
<th>UAS</th>
<th>Oracle (best possible src)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td>Telugu</td>
<td>66.7</td>
<td>✓</td>
</tr>
<tr>
<td>Czech</td>
<td>Slovak</td>
<td>65.8</td>
<td>✓</td>
</tr>
<tr>
<td>Danish</td>
<td>Slovenian</td>
<td>42.1</td>
<td>+13.3 English</td>
</tr>
<tr>
<td>German</td>
<td>English</td>
<td>56.8</td>
<td>✓</td>
</tr>
<tr>
<td>Slovak</td>
<td>Slovenian</td>
<td>58.4</td>
<td>+ 3.3 Czech</td>
</tr>
<tr>
<td>Tamil</td>
<td>Turkish</td>
<td>31.1</td>
<td>+22.4 Hindi</td>
</tr>
</tbody>
</table>
Average over 18 test TBs

- Oracle src identified by $KL_{cpos3}$ in 8 cases
- Average best $KL_{cpos3} = 0.7$
  - $< 0.7 \rightarrow 7x$ oracle, 2x competitive, 3x bad
  - $> 0.7 \rightarrow 1x$ oracle, 5x bad
Parse tree combination

src 1: 

src 2: 

src 3: 

tgt:
Parse tree combination

src 1: #root VERB PREP NOUN

+ src 2: #root VERB PREP NOUN

+ src 3: #root VERB PREP NOUN

= tgt: #root VERB PREP NOUN
Weighted parse tree combination

\[ KL_{cpos3}^{-4} : \]

\[ x 1.9 \]

\[ x 1.7 \]

\[ x 0.5 \]

\[ = \text{tgt:} \]

\[ \frac{4.1}{2.4} \]

\[ \frac{2.2}{1.9} \]

\[ \frac{1.7}{NOUN} \]
Weighted parse tree combination

\[ KL_{cpos3}^{-4} : \]

\[
\begin{align*}
\text{src 1:} & \quad \# \text{root} & \text{VERB} & \text{PREP} & \text{NOUN} & \times 1.9 \\
+ \text{src 2:} & \quad \# \text{root} & \text{VERB} & \text{PREP} & \text{NOUN} & \times 1.7 \\
+ \text{src 3:} & \quad \# \text{root} & \text{VERB} & \text{PREP} & \text{NOUN} & \times 0.5 \\
\text{tgt:} & \quad \# \text{root} & \text{VERB} & \text{PREP} & \text{NOUN} & \\
\end{align*}
\]
Average over 18 test TBs

- Treebank concatenation: 45%
- Single-source selection: 49%
- Unweighted combination: 48%
- Weighted combination: 53%
- Oracle src: 56%
Conclusion

- $KL_{cpos 3}$ language similarity measure
  - for delexicalized parser transfer
- single-source transfer
  - source treebank selection (44% success)
- multi-source transfer (tree combination)
  - source treebank weighting with $KL_{cpos 3}^{-4}$
  - +3.9% over single-source transfer
  - +4.5% over unweighted tree combination
  - +8.0% over treebank concatenation
Thank you for your attention

Rudolf Rosa, Zdeněk Žabokrtský
{rosa,zabokrtsky}@ufal.mff.cuni.cz

$KL_{cpos}^3$

a Language Similarity Measure for Delexicalized Parser Transfer

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

http://ufal.mff.cuni.cz/rudolf-rosa/