Multi-source Cross-lingual Delexicalized Parser Transfer: Prague or Stanford?

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

Depling, Uppsala, 26 August 2015
Introduction

- **Task**: cross-lingual parser transfer
  - train a parser on treebank for language A
  - use the parser to parse text in language B
  - HamleDT 2.0 treebank collection (30 languages)

- **Question**: Prague or Stanford annotation style?
  - focus on adposition annotation (= prep., postp., ...)
  - Prague: ADP as head  x  Stanford: ADP as leaf
Motivation

- Prague style generally better for parsing
  - full Prague +5% abs. over full Stanford (avg UAS)
  - Prague ADP +0.8% over Stanford ADP – lexicalized
    - delexicalized: +0.2% (usually significant, but weaker)
Motivation

- Prague style generally better for parsing
  - full Prague +5% abs. over full Stanford (avg UAS)
  - Prague ADP +0.8% over Stanford ADP – lexicalized
    - delexicalized: +0.2% (usually significant, but weaker)

- Stanford style more cross-lingually consistent
  - might be beneficial for cross-lingual parsing
General Approach

- use Prague-style HamleDT treebanks
- perform automatic conversions between Prague and Stanford ADP annotation style
- vary annotation style used at:
  - parser training
  - parser transfer
  - final output
- evaluate with UAS (unlabelled attachment score)
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 2.0: 30 langs)
    - unannotated data (here: POS tagged)
(Lexicalized) MSTParser

Rudolf Rosa: Multi-source Cross-lingual Delexicalized Parser Transfer: Prague or Stanford?
(Lexicalized) MSTParser

Rudolf Rosa: Multi-source Cross-lingual Delexicalized Parser Transfer: Prague or Stanford?
Delexicalized MSTParser

#root → NOUN → VERB → NOUN

7.4 → -4.3 → 9.7 → 10.8

16.1 → -2.4 → 2.2

16.1 → 9.7 → 10.8
Delexicalized parser transfer

- **Single-source**
  - train a delexicalized parser on a *source* language treebank (e.g. Czech – PDT)
  - use it to parse *target* language text (e.g. Slovak)
    - need a POS tagger for the target language

- **Multi-source**
  - a set of source treebanks, train a parser on each
  - parse target text by each of the parsers
  - combine their outputs
Parse tree combination

src 1:
- #root
- VERB
- ADP
- NOUN

+ src 2:
- #root
- VERB
- ADP
- NOUN

+ src 3:
- #root
- VERB
- ADP
- NOUN

= tgt:
- #root
- VERB
- ADP
- NOUN
Parse tree combination

src 1:

+ src 2:

+ src 3:

= tgt:
Experiments

- vary ADP annotation styles
  - Prague style (“P”), Stanford style (“S”)
  - (other phenomena = always Prague style)
- +/- convert treebanks to style X
  - train delexicalized parsers
  - parse the target text by the parsers
- +/- convert parser outputs to style Y
  - combine the parse trees
- setups denoted “X/Y” (e.g. “P/S”)
Smaller Source Treebank Subsets

- sources for training: smaller groups of treebanks grouped by frequency of ADP
  - high: Spanish (15%), Hindi (19%), Japanese (19%)
  - medium: English (8%), Czech (9%), Swedish (9%)
  - low: Basque (0%), Hungarian (1%), Anc. Greek (4%)
  - mix: Hungarian (1%), Swedish (9%), Spanish (15%)
  - all 9

- targets for testing: the remaining 21 treebanks
Smaller Source Treebank Subsets

<table>
<thead>
<tr>
<th>Subset</th>
<th>P/P</th>
<th>P+S/P</th>
<th>S/S</th>
<th>P+S/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Med</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mix</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- adposition annotation style
  - Prague (ADP=head) vs Stanford (ADP=leaf)
- multisource crosslingual parser transfer
- Stanford style seems better
  - even more so for small and diverse treebank sets
- best results:
  - use both Prague and Stanford style for training
  - convert to Stanford style for combination and output
Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Multi-source Cross-lingual Delexicalized Parser Transfer: Prague or Stanford?

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

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