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

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

-22.3  48.7  -5.2  16.4  20.7  -12.4  7.3
(Lexicalized) MSTParser

Rudolf NOUN likes VERB trains NOUN

#root

Rudolf NOUN
likes VERB
trains NOUN

#root
Rudolf NOUN
likes VERB
trains NOUN

48.7
16.4
20.7
7.3
13.8
22.3
-5.2
32.5
-12.4
Delexicalized MSTParser
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 → train a parser → 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
\[ KL_{cpos^3} (tgt, src) = \sum_{\forall cpos^3 \in tgt} f_{tgt}(cpos^3) \cdot \log \left( \frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)} \right) \]

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

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

| Target lang. | KL\textsubscript{cpos 3} selected src | UAS | Oracle (best possible src) |
|--------------|--------------------------------------|-----|---------------------------|
| Bengali      | Telugu                              | 66.7| ✓                         |
| Czech        | Slovak                              | 65.8| ✓                         |
| Danish       | Slovenian                            | 42.1| +13.3 English             |
| German       | English                             | 56.8| ✓                         |
| Slovak       | Slovenian                            | 58.4| + 3.3 Czech               |
| Tamil        | Turkish                             | 31.1| +22.4 Hindi              |
Average over 18 test TBs

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

src 1: 

VERB
PREP
NOUN

+ src 2: 

VERB
PREP
NOUN

+ src 3: 

VERB
PREP
NOUN

= tgt: 

VERB
PREP
NOUN
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

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

\[ x 1.9 \]

\[ x 1.7 \]

\[ x 0.5 \]

\[ = \text{tgt:} \]

\[ 4.1 \]

\[ 2.4 \]

\[ 2.2 \]

\[ 1.9 \]

\[ 1.7 \]

\[ \text{src 1:} \]

\[ \text{src 2:} \]

\[ \text{src 3:} \]
Weighted parse tree combination

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

\begin{align*}
\text{src 1:} & \quad \#root \quad \text{VERB} \quad \text{PREP} \quad \text{NOUN} \\
+ \text{src 2:} & \quad \#root \quad \text{VERB} \quad \text{PREP} \quad \text{NOUN} \\
+ \text{src 3:} & \quad \#root \quad \text{VERB} \quad \text{PREP} \quad \text{NOUN} \\
= \text{tgt:} & \quad \#root \quad \text{VERB} \quad \text{PREP} \quad \text{NOUN}
\end{align*}

\begin{align*}
\text{src 1:} & \quad 4.1 \\
\text{src 2:} & \quad 2.4 \\
\text{src 3:} & \quad 2.2 \\
\text{tgt:} & \quad 1.9 \\
& \quad 1.7
\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/