Parsing Natural Language Sentences by Semi-supervised Methods

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

Dissertation interim report defence, Praha, 15 June 2015
Outline

- Intro and motivation
- MSTParser and its delexicalization
- Delexicalized parser transfer
  - single-source
  - multi-source
    - tree combination
    - model interpolation
- Choice of parsing annotation style
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

#root

Rudolf
NOUN

likes
VERB

trains
NOUN

22.3

48.7

13.8

-5.2

32.5

16.4

20.7

-12.4

7.3
(Lexicalized) MSTParser
Delexicalized MSTParser

Diagram:

- #root
- NOUN
- VERB
- NOUN

Connectors:
- #root to NOUN: 7.4
- NOUN to VERB: 16.1
- VERB to NOUN: -4.3
- NOUN to #root: 10.8
- VERB to NOUN: 2.2
- NOUN to VERB: 5.2
- VERB to NOUN: -1.4
- NOUN to VERB: 9.7
- VERB to NOUN: -2.4
Single-source delex parser transfer

- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
Single-source delex parser transfer

- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
- how to choose the src lang for a tgt lang?
Single-source delex parser transfer

- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
- how to choose the src lang for a tgt lang?
  - src should be as similar to tgt as possible
Single-source delex parser transfer

- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
- how to choose the src lang for a tgt lang?
  - src should be as similar to tgt as possible
  - $KL_{cpos3}$: POS trigram distribution in tagged corpora
Multi-source: 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
Multi-source: tree combination

src 1:
  #root  \rightarrow  VERB  \rightarrow  PREP  \rightarrow  NOUN

+ src 2:
  #root  \rightarrow  VERB  \rightarrow  PREP  \rightarrow  NOUN

+ src 3:
  #root  \rightarrow  VERB  \rightarrow  PREP  \rightarrow  NOUN

= tgt:
  #root  \rightarrow  VERB  \rightarrow  PREP  \rightarrow  NOUN
Multi-source: tree combination

KL_{cpos3}^{-4}:

src 1: #root | VERB | PREP | NOUN

+ src 2: #root | VERB | PREP | NOUN

+ src 3: #root | VERB | PREP | NOUN

= tgt: #root | VERB | PREP | NOUN

4.1  2.4  1.9

x 0.5  x 1.7  x 1.9
Multi-source: 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

KL_{clos3}^{-4}:
- x 1.9
- x 1.7
- x 0.5

KL_{clos3}^{-4}:
- x 1.9
- x 1.7
- x 0.5
Multi-source: model interpolation

src 1:

#root 17.4 VERB 14.3 PREP 9.7 NOUN

src 2:

#root 12.5 VERB 3.4 PREP 15.7 NOUN
Multi-source: model interpolation

tgt (igsaw): #root 29.9

VERB 17.7

PREP 16.8

NOUN 25.4

-1.1
Multi-source: model interpolation

= tgt:

#root -> VERB (29.9) -> PREP (17.7) -> NOUN (25.4)

-1.1  2.7  16.8  12.9  7.8  3.5
Evaluation on HamleDT (30 langs)

- **Single-source**
  - Weighted: 50%
  - Unweighted: 46%

- **Tree combination**
  - Weighted: 53%
  - Unweighted: 49%

- **Model interpolation**
  - Weighted: 53%
  - Unweighted: 47%

UAS scores range from 42% to 56%.
Annotation style (multi-source)

- Prague 57% UAS, Stanford 49% UAS
- Prague 57% UAS, Stanford 49% UAS
  - Prague better as base; Stanford adpositions?
Annotation style (multi-source)

- Prague 57% UAS, Stanford 49% UAS
  - Prague better as base; Stanford adpositions?

- interesting results when combining both adposition annotation styles (+0.39% UAS)
  - solitary langs: big improvement (et +3%, fa +2%)
Conclusion

- Delexicalized parser transfer
  - single-source
  - multi-source: tree combination, model interpolation
  - $KL_{cpos3}$: lang similarity for src selection/weighting

- Annotation style for parsing
  - Prague better than Stanford
  - Stanford adpositions good for cross-lingual transfer
Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Parsing Natural Language Sentences by Semi-supervised Methods

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

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