Recent Advances in Dependency Parsing

Tutorial, EACL, April 27th, 2014

Ryan McDonald\textsuperscript{1} Joakim Nivre\textsuperscript{2}

\textsuperscript{1}Google Inc., USA/UK
E-mail: ryanmcd@google.com

\textsuperscript{2}Uppsala University, Sweden
E-mail: joakim.nivre@lingfil.uu.se
Overview of the Tutorial

- Introduction to Dependency Parsing (Joakim)
- Graph-based parsing post-2008 (Ryan)
- Transition-based parsing post-2008 (Joakim)
- **Summary and final thoughts** (Ryan)
Topics Not Discussed
Unsupervised Learning

Learn models from unlabeled data only

- Dependency Model with Valency (DMV) [Klein and Manning 2004]
- Adding in prior knowledge/constraints
  - Sparsity over head-modifier POS combinations [Gillenwater et al. 2010]
  - Universal linguistic knowledge [Naseem et al. 2010]
Semi-Supervised Learning

Learn models from labeled and unlabeled data

- Co-training [Sagae and Tsujii 2007]
- Tri-training [Søgaard and Rishøj 2010a]
- Up/self-training [Petrov et al. 2010]
- Structured conditional model [Suzuki et al. 2009]
- Web-derived features [Bansal and Klein 2011]
- Dependency language model [Chen et al. 2012]
- Meta-features [Chen et al. 2013]
Cross-Lingual Learning

Learn models from foreign or parallel language resources

- **Projection** [Hwa et al. 2005, Ganchev et al. 2009]
- **Delexicalized transfer** [Zeman and Resnik 2008]
- **Multi-source training** [McDonald et al. 2011, Cohen et al. 2011]
- **Linguistic priors** [Naseem et al. 2012, Täckström et al. 2013]
- **Cross-lingual resources** [Täckström et al. 2012, Durrett et al. 2012]

- **Significantly higher accuracies than unsupervised learning**
  [McDonald et al. 2011]
Learning with Approximate Search

- Learning algorithms assume exact search
- Pruning, beam search and other approximations break this
- Huang et al. [2012]: principled method for perceptron with approximate search
- Transition-based parsing [Huang et al. 2012]
  - English UAS: 92.1 → 92.2
  - Speeds up training by factor of 3
- Graph-based parsing [Zhang et al. 2013]
  - English LAS: 92.92 → 93.64
  - English LAS: 90.35 → 91.28
  - Bottom-up parsing has larger search space
Domain Adaptation

- Dependency parsers are subject to domain shift
  - WSJ LAS $\rightarrow$ QTB LAS: 86.4 $\rightarrow$ 67.0 [Petrov et al. 2010]
  - WSJ LAS $\rightarrow$ Web LAS: 91.5 $\rightarrow$ 83.4 [Petrov and McDonald 2012]

- Ensembles and self-training [Sagae and Tsujii 2007]
- Datapoint selection [Kawahara and Uchimoto 2008]
- Grammars + statistical parsers
  [Zhang and Wang 2009, Petrov et al. 2010]
- Tri-training [Søgaard and Rishøj 2010b]
- Training with domain specific loss functions [Hall et al. 2011]
- Shared tasks
  [Nivre et al. 2007, Dell’Orletta et al. 2011, Petrov and McDonald 2012]
Introduction

Parsing General Graphs

- Tree constraint often just a computational convenience
- Conceptually, dependency graphs don’t need to be trees
- Directed arcs between words can encode:
  - Raising and control structures
  - More direct co-ordination structures
  - Traces and wh-movement
  - ...
- Such structures are more semantic in nature

- McDonald and Pereira [2006] parse DAGs with approximate graph-based inference
- Sagae and Tsujii [2008] extend transition-based system to parse DAGs
- Many other solutions in CoNLL 2008 and 2009 shared tasks
  [Surdeanu et al. 2008, Hajič et al. 2009]
Using Phrase-Structure Parsers

- Lexicalized phrase-structure parsers produce dependencies
  [Collins 1999, Charniak 2000]
- Phrase-structure parsers can also produce dependencies via post-processing [Cer et al. 2010]
  - Latter is more accurate [McDonald et al. 2005]
  - Bias in post-processing [Petrov and McDonald 2012]
- Dependency to phrase-structure treebank conversion
  [Collins et al. 1999]
- Coppola and Steedman [2013]: cube-pruned phrase-structure parser with dependency features
  - Highest reported En scores for both phrase-structure and dependency evaluations
  - Caveat: conversion heuristics to generate dependencies
  - Can combine with cube-pruned dependency parsers
Improved Evaluations

- LAS/UAS evaluations are useful, but
  - Treat all errors as equal
  - Don’t say anything about downstream performance
  - Only allow comparisons on single annotation scheme

- Targeted dependency evaluations
  - Long-distance and/or implicit dependencies
    [Rimell et al. 2009, Nivre et al. 2010]

- Downstream task evals [Miyao et al. 2008]
  - Training to optimize task specific evals [Hall et al. 2011]

- Annotation-scheme independent evaluations [Tsarfaty et al. 2011]
Final Thoughts
Where do we stand?

**Evaluated on overlapping 9 languages in studies**

---

**Las: 83.8 v. 83.6**
[McDonald & Nivre 2007]

**Las: 85.8 v. 85.5**
[Zhang et al. 2013]
Where do we stand?

▶ 2008
  ▶ Graph-based and transition-based have near identical accuracies [Buchholz and Marsi 2006]
  ▶ But very different errors [McDonald and Nivre 2007]
  ▶ And errors can be correlated with model properties

▶ 2014
  ▶ 2008+: Attempts to address model short-comings
  ▶ Models have converged (structured prediction, rich features, heuristic inference)
  ▶ Accuracies again nearly identical [Zhang et al. 2013]
Where do we stand?

- **2008**
  - Graph-based and transition-based have near identical accuracies [Buchholz and Marsi 2006]
  - But very different errors [McDonald and Nivre 2007]
  - And errors can be correlated with model properties

- **2014**
  - 2008+: Attempts to address model short-comings
  - Models have converged (structured prediction, rich features, heuristic inference)
  - Accuracies again nearly identical [Zhang et al. 2013]

Do graph-based and transition-based parsers still make different kinds of errors?
Where do we stand?

- **2008**
  - Graph-based and transition-based have near identical accuracies [Buchholz and Marsi 2006]
  - But very different errors [McDonald and Nivre 2007]
  - And errors can be correlated with model properties

- **2014**
  - 2008+: Attempts to address model short-comings
  - Models have converged (structured prediction, rich features, heuristic inference)
  - Accuracies again nearly identical [Zhang et al. 2013]

Are the remaining model differences meaningful? Formal/statistical power? Empirically?
References and Further Reading


- Daniel M Cer, Marie-Catherine De Marneffe, Daniel Jurafsky, and Christopher D Manning. 2010. Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *LREC*.


- Wenliang Chen, Min Zhang, and Haizhou Li. 2012. Utilizing dependency language models for graph-based dependency parsing models. In *Proceedings of the 50th*
References and Further Reading


References and Further Reading


References and Further Reading
References and Further Reading


