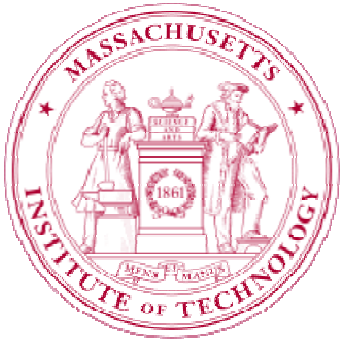
Neutralizing Linguistically Problematic Annotations in Unsupervised Dependency Parsing Evaluation

Roy Schwartz[1], Omri Abend[1],
Roi Reichart[2] and Ari Rappoport[1]

[1]The Hebrew University, [2]MIT

In proceedings of ACL 2011

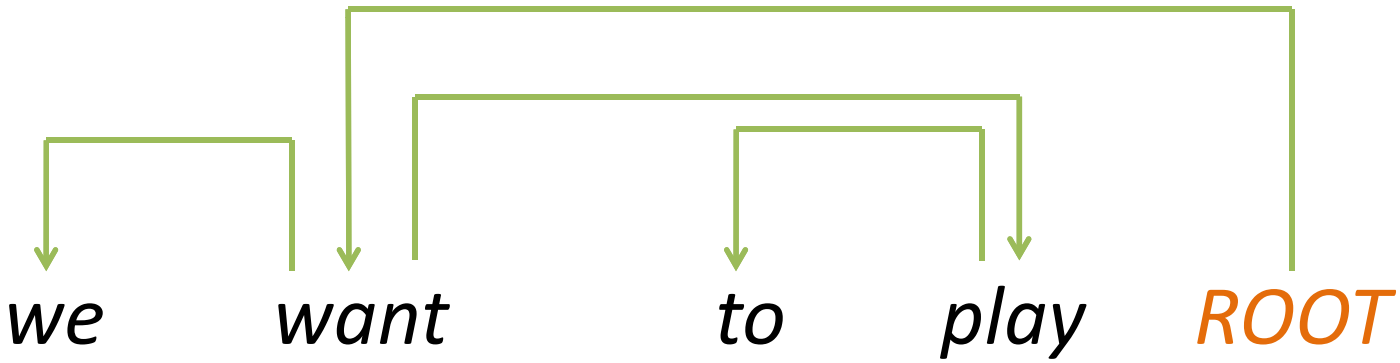


Outline

- Introduction
- Problematic Gold Standard Annotation
- Sensitivity to the Annotation of Problematic Structures
- A Possible Solution – Undirected Evaluation
- A Novel Evaluation Measure

Introduction

Dependency Parsing



Introduction
Related Work

- Supervised Dependency Parsing
  - McDonald et al., 2005
  - Nivre et al., 2006
  - Smith and Eisner, 2008
  - Zhang and Clark, 2008
  - Martins et al., 2009
  - Goldberg and Elhadad, 2010
  - *inter alia*

- Unsupervised Dependency Parsing (unlabeled)
  - Klein and Manning, 2004
  - Cohen and Smith, 2009
  - Headden et al., 2009
  - Blunsom and Cohn, 2010
  - Spitkovsky et al., 2010
  - *inter alia*

Introduction

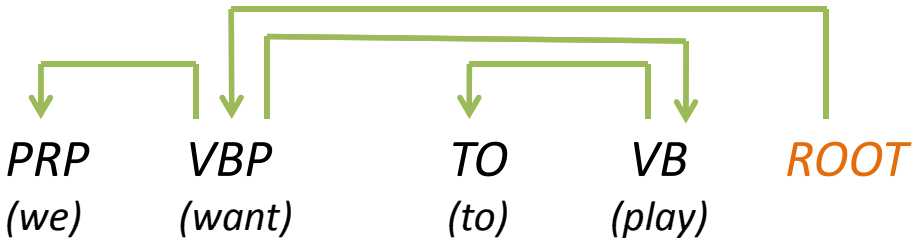
Unsupervised Dependency Parsing Evaluation

- Evaluation performed against a gold standard
- Standard Measure – *Attachment Score*
  - Ratio of correct *directed* edges
- A single score (no precision/recall)

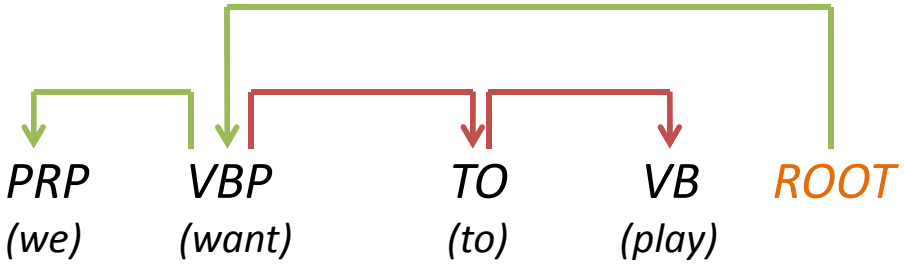
Introduction

Unsupervised Dependency Parsing Evaluation

- Example
  - *Gold Std:*



  - *Score: 2/4*



Problematic Gold Standard Annotation

- The gold standard annotation of some structures is **Linguistically Problematic**
  - I.e., *not under consensus*
- Examples
  - Infinitive Verbs

(Collins, 1999)

to ⇄ play

(Bosco and Lombardo, 2004)

  - Prepositional Phrases

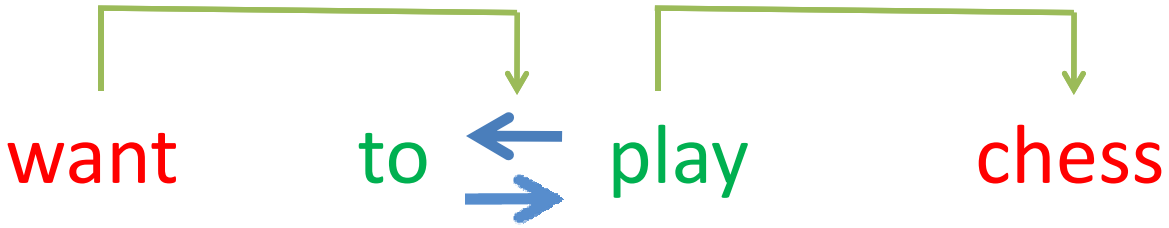
(Johansson and Nugues, 2007)

in ⇄ Rome

(Yamada and Matsumoto, 2003)

Problematic Gold Standard Annotation

- Great majority of the problematic structures are local
  - Confined to 2–3 words only
  - Often, alternative annotations differ in the direction of some edge
  - The controversy only relates to the **internal** structure

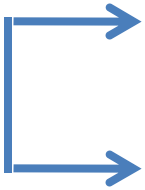


- These structures are also very frequent
  - 42.9% of the tokens in PTB WSJ participate in at least one problematic structure

Problematic Gold Standard Annotation

- Gold standard in English (and other languages) – converted from constituency parsing using head percolation rules

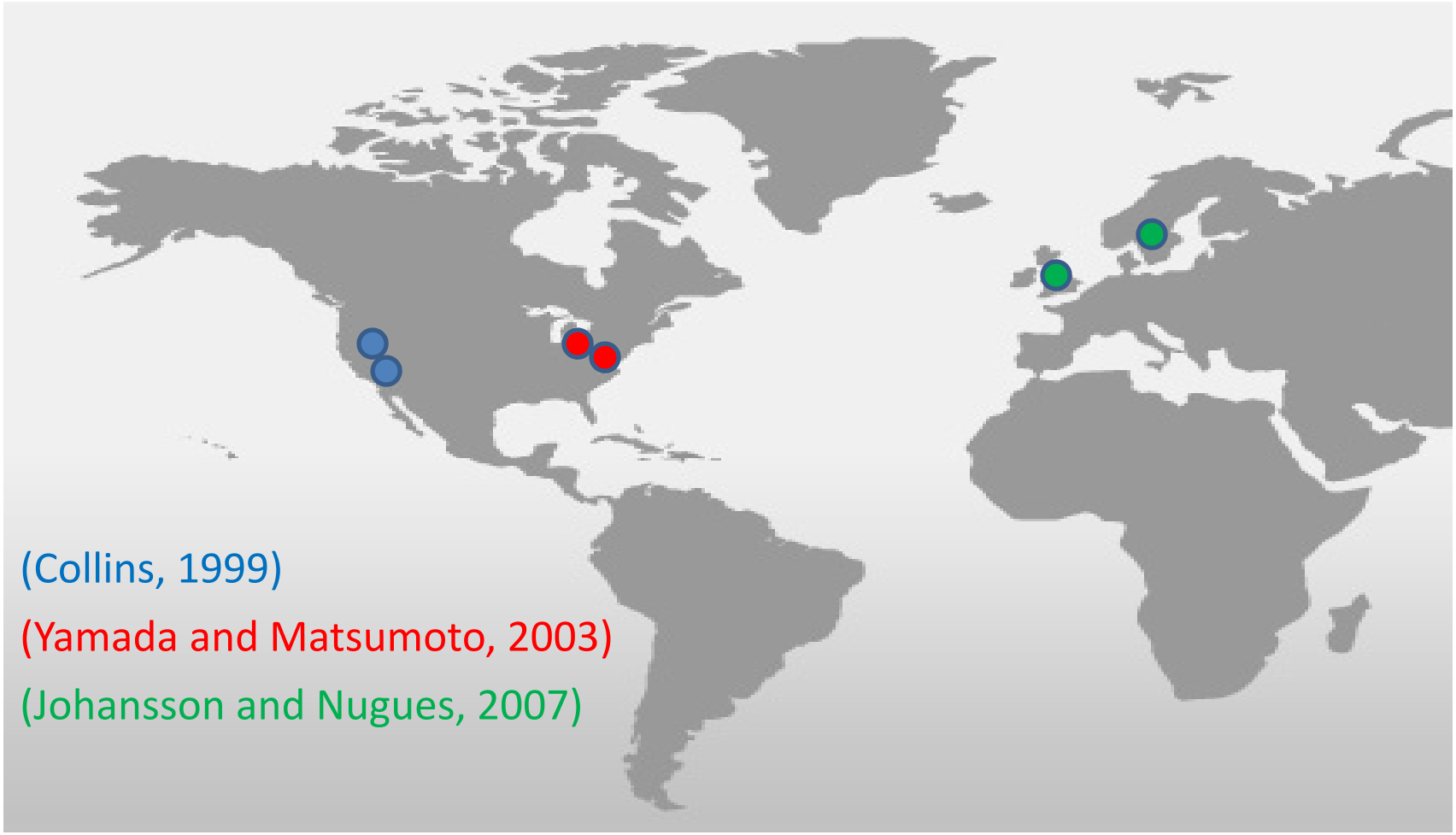
- At least **three substantially different** conversion schemes are currently in use for *the same task*



1. Collins head rules (Collins, 1999)
   - Used in e.g., (Berg-Kirkpatrick et al., 2010; Spitkovsky et al., 2010)
2. Conversion rules of (Yamada and Matsumoto, 2003)
   - Used in e.g., (Cohen and Smith, 2009; Gillenwater et al., 2010)
3. Conversion rules of (Johansson and Nugues, 2007)
   - Used in e.g., the CoNLL shared task 2007, (Blunsom and Cohn, 2010)

14.4% Diff.

Problematic Gold Standard Annotation



Problematic

Structures

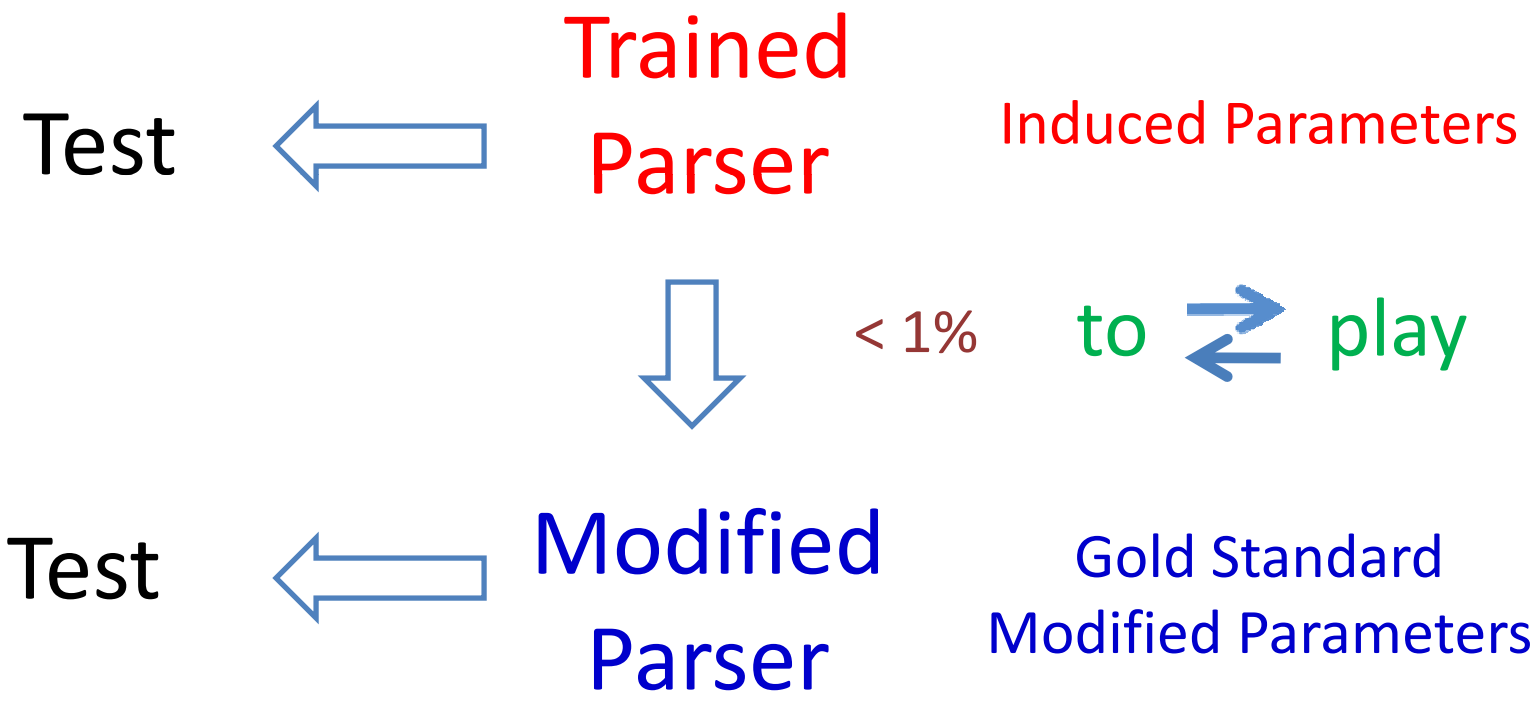
Very *Frequent*

3 *Substantially Different*

Gold Standards

Evaluation Problem

Sensitivity to the Annotation of Problematic Structures



Sensitivity to the Annotation of Problematic Structures

Model	Original	Modified	Modified - Original
km04	34.3	43.6	**9.3**
cs09	39.7	54.4	**14.7**
saj10	41.3	54	**12.7**

- *km04* – Klein and Manning, 2004
- *cs09* – Cohen and Smith, 2009
- *saj10* – Spitkovsky et al., 2010

Current evaluation

*does **not** always*

reflect parser quality

A Possible Solution

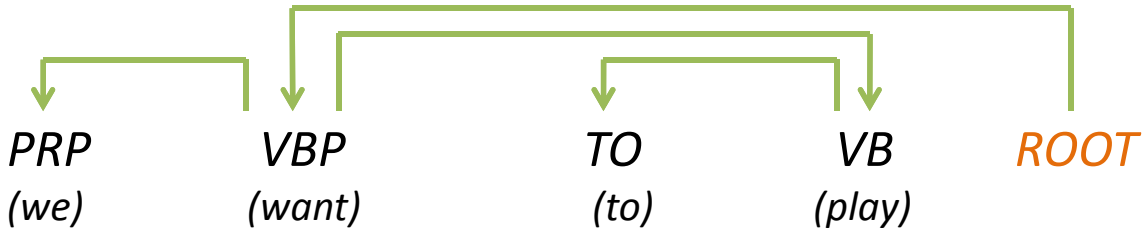
Undirected Evaluation

- **Required** – a measure indifferent to alternative annotations of problematic structures
- **Recall** – most alternative annotations differ only in the direction of some edge
- **A possible solution** – a measure indifferent to edge directions
- How about *undirected evaluation*?

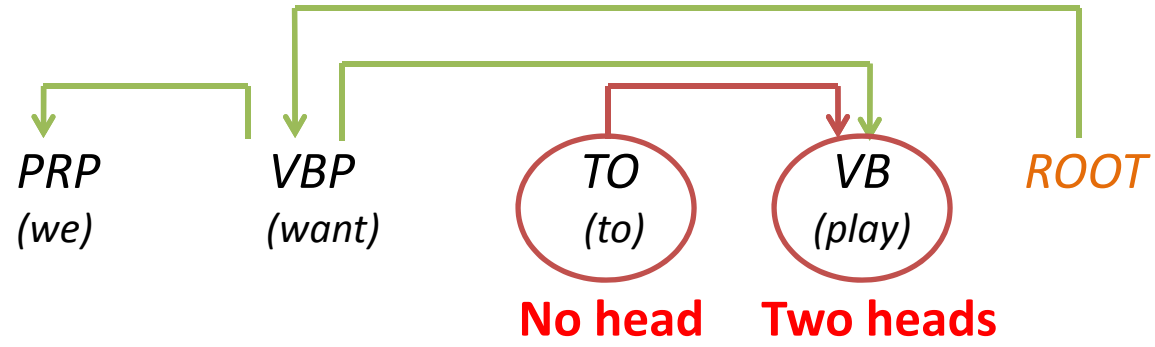
A Possible Solution

Undirected Evaluation

- Gold standard:



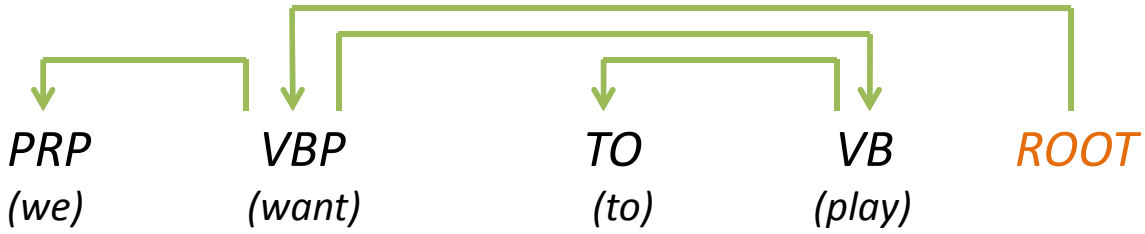
- Induced parse, with a flipped edge



A Possible Solution

Undirected Evaluation

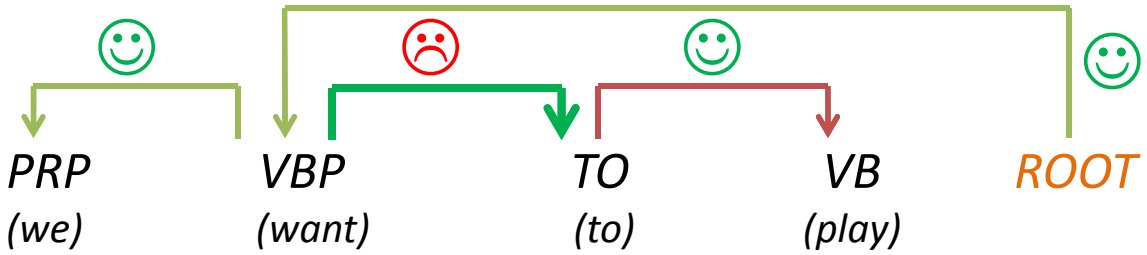
- Gold standard:



- Induced parse, with a flipped edge

3/4 (75%)

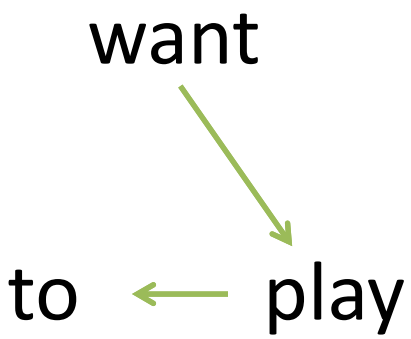
This is the minimal undirected score modification!



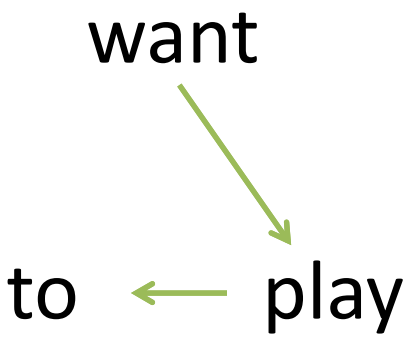
The Neutral Edge Direction (NED) Measure

- Undirected accuracy is *not indifferent* to edge flipping

- We will now present a measure that is – *Neutral Edge Direction* (*NED*)
  - A simple extension of the undirected evaluation measure
  - Ignores edge direction flips

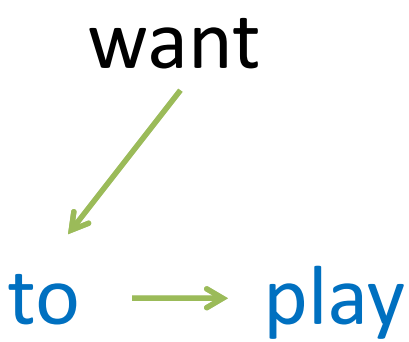


Gold Standard



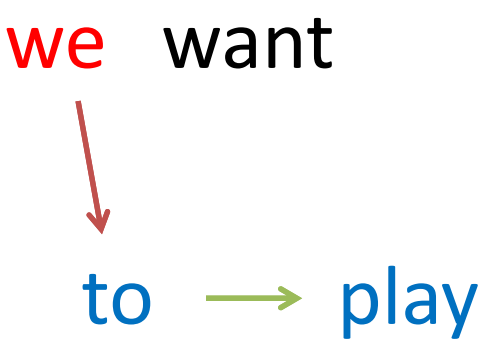
Induced parse I
(agrees with gold std.)

•correct NED attachment



Induced parse II
(linguistically plausible)

•correct NED attachment



Induced parse III
(linguistically implausible)

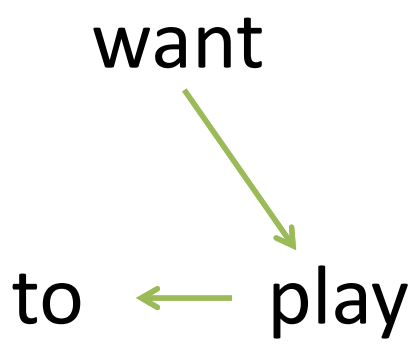
•NED error

The NED Measure

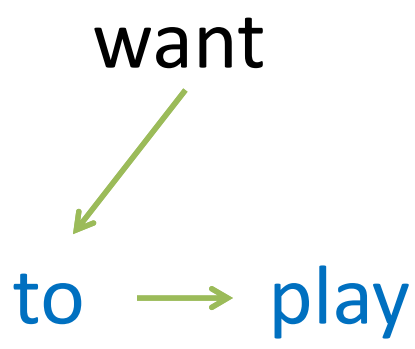
- Therefore, NED is defined as follows:
  - X is a correct parent of Y if:
    - X is Y's gold parent **or**
    - X is Y's gold child **or**
    - X is Y's gold grandparent

} Attachment

} Undirected



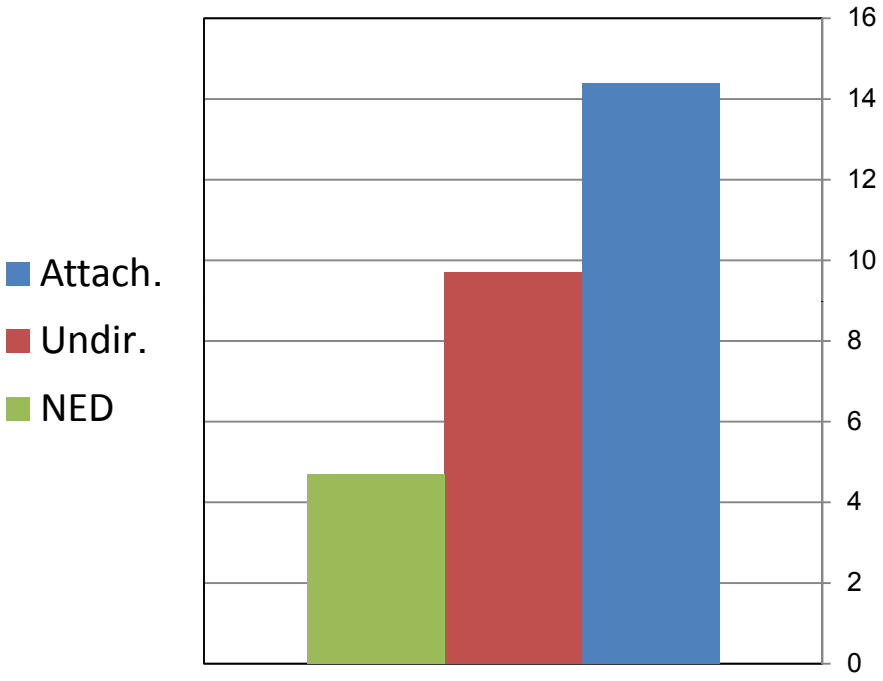
Gold Standard



linguistically plausible parse

NED Experiments

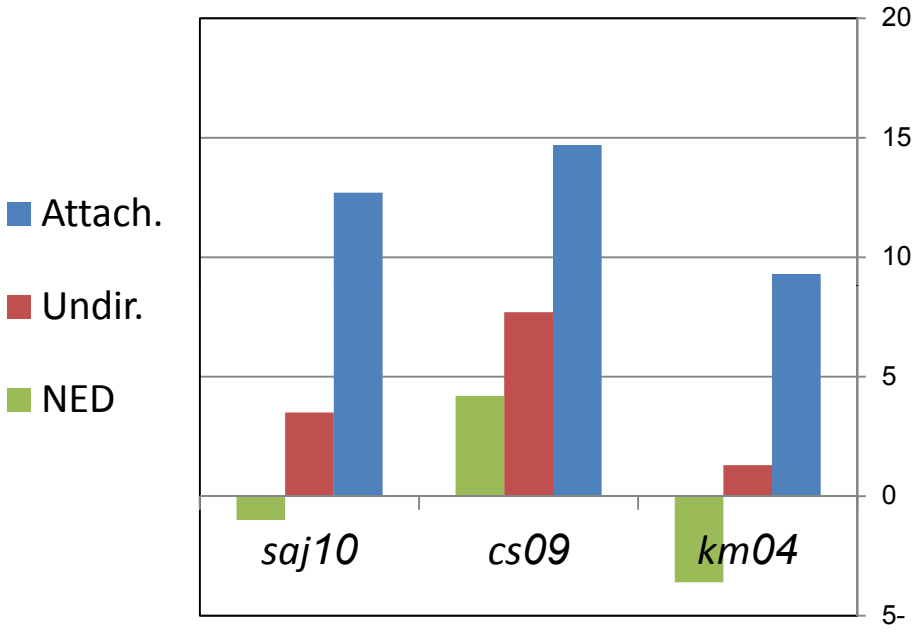
Difference Between Gold Standards



- *NED* substantially reduces the difference between alternative gold standards

NED Experiments

Sensitivity to Parameter modification



- *NED* substantially reduces the difference between parameter sets
- The sign of the NED difference is predictable (see paper)

Discussion

- Unsupervised parsers train on *plain text*
  - Choosing the "wrong" (*plausible*) annotation should not be considered an error
  - Use NED!

- Supervised parsers train on *labeled data*
  - They get the correct annotation as training input

- Neverthless, NED can be used to *better understand* the type of errors performed by *supervised* parsers
  - Better suited than using undirected evaluation measure

Future Work

- Find a more fine-grained measure
  - *Evaluating Dependency Parsing: Robust and Heuristics-Free Cross-Annotation Evaluation* (Tsarfaty et al., to appear in *EMNLP 2011*)

- Resolve conflicts in annotation level

Summary

- Problems in the evaluation of unsupervised parsers
  - **Gold Standards** – 3 used (~15% difference between them)
  - **Current Parsers** – very sensitive to alternative (plausible) annotations. Minor modifications result in ~9–15% performance "gain"
  - **Undirected Evaluation** – does not solve this problem

- Neutral Edge Direction (NED) measure
  - Simple and intuitive
  - Reduces difference between different gold standards to ~5%
  - Reduces undesired performance "gain" (~1–4%)
  - Still *indicative* of *quality difference*
    - See more experiments demonstrating NED's validity (see paper)

Take–Home Message

- We suggest reporting NED results along with the commonly used attachment score

http://www.cs.huji.ac.il/~roys02/software/ned.html

Many thanks to

- Shay Cohen
- Valentin I. Spitkovsky
- Jennifer Gillenwater
- Taylor Berg-Kirkpatrick
- Phil Blunsom