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

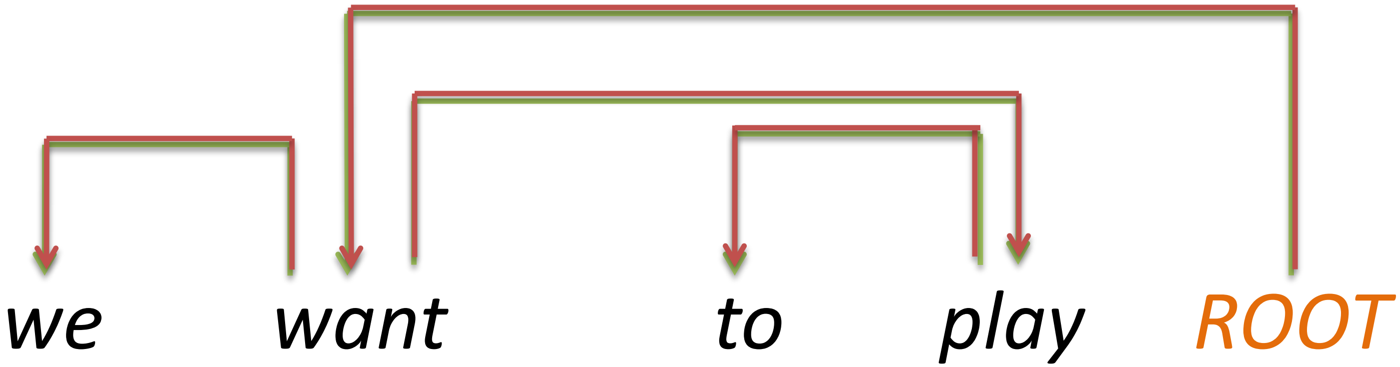
ISCOL 2011

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



we *want* *to* *play* *ROOT*

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

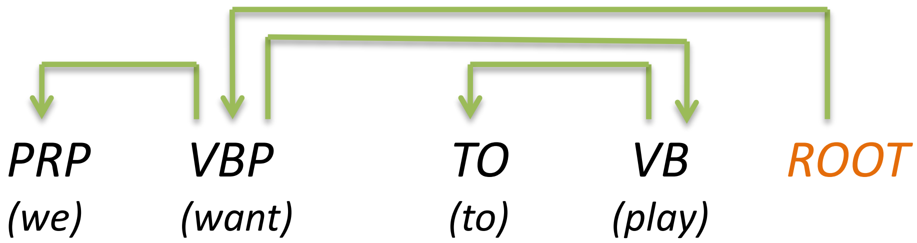
- Standard Measure – *Attachment Score*
  - Ratio of correct *directed* edges

- A single score (no precision/recall)

Introduction

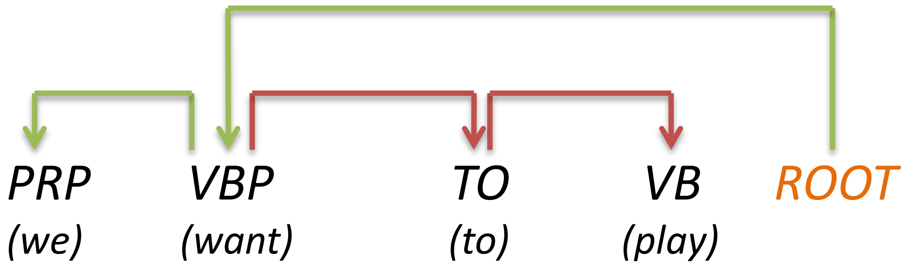
Unsupervised Dependency Parsing Evaluation

- Example
  - *Gold Std:*



PRP *(we)* *VBP* *(want)* *TO* *(to)* *VB* *(play)* *ROOT*

  - *Score: 2/4*

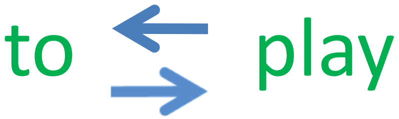


PRP *(we)* *VBP* *(want)* *TO* *(to)* *VB* *(play)* *ROOT*

Problematic Gold Standard Annotation

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

(Collins, 1999)



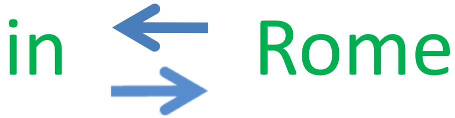
to

play

(Bosco and Lombardo, 2004)

  - Prepositional Phrases

(Johansson and Nugues, 2007)



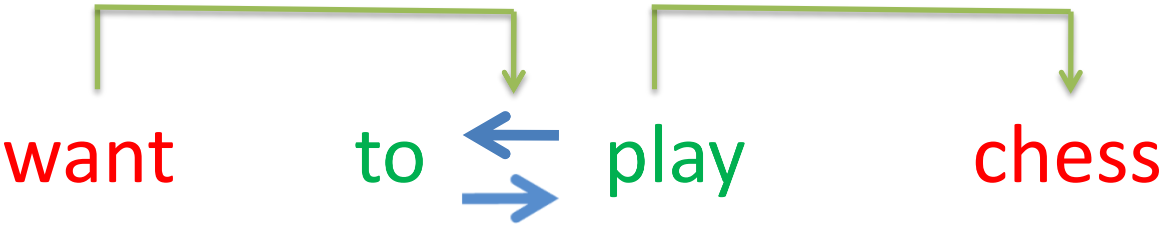
in

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

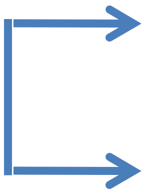


- These structures are also very frequent
  - 42.9% of the tokens in PTB WSJ participate in at least one problematic structure

Problematic Gold Standard Annotation

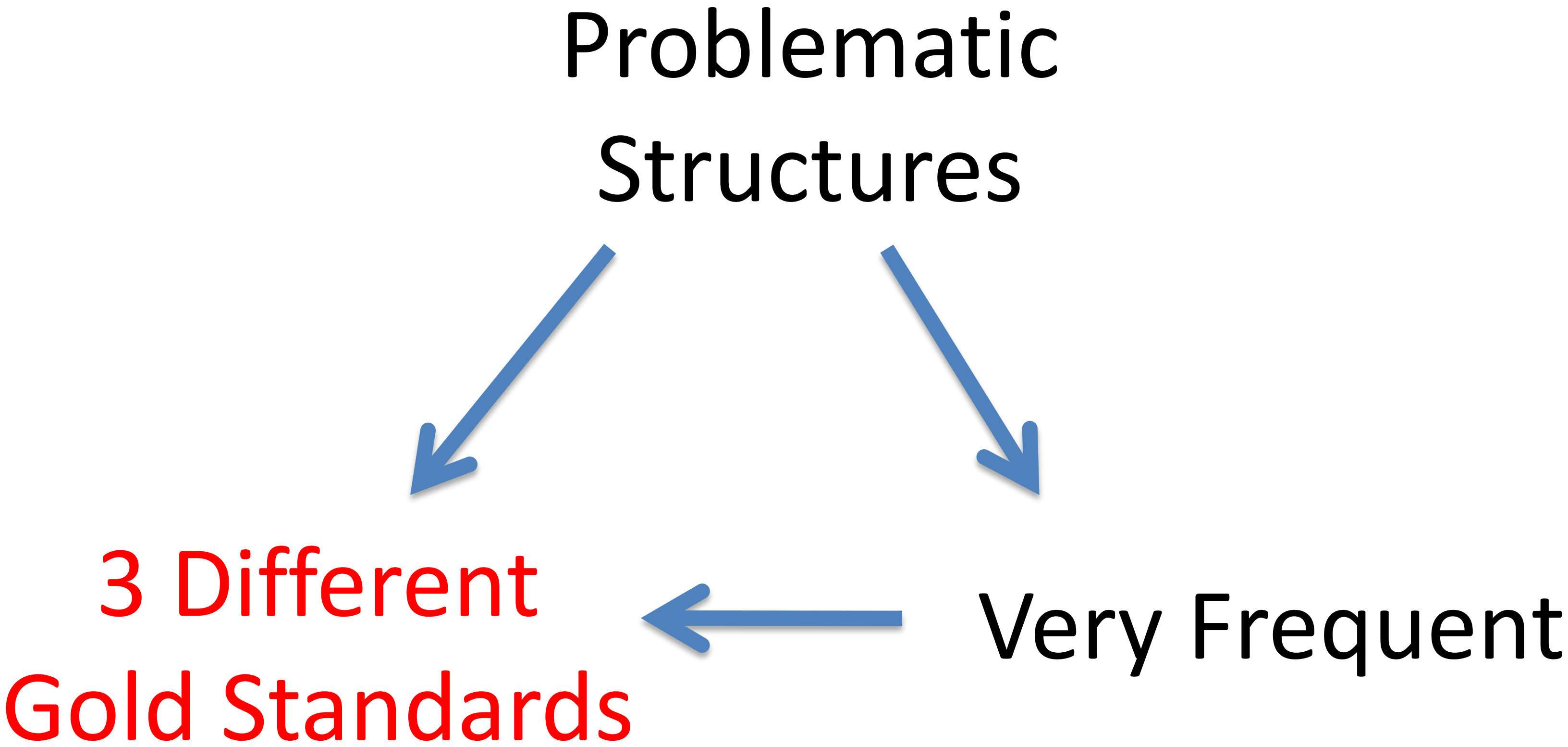
- Gold standard in English (and other languages) – converted from constituency parsing using head percolation rules

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



14.4% Diff.

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)

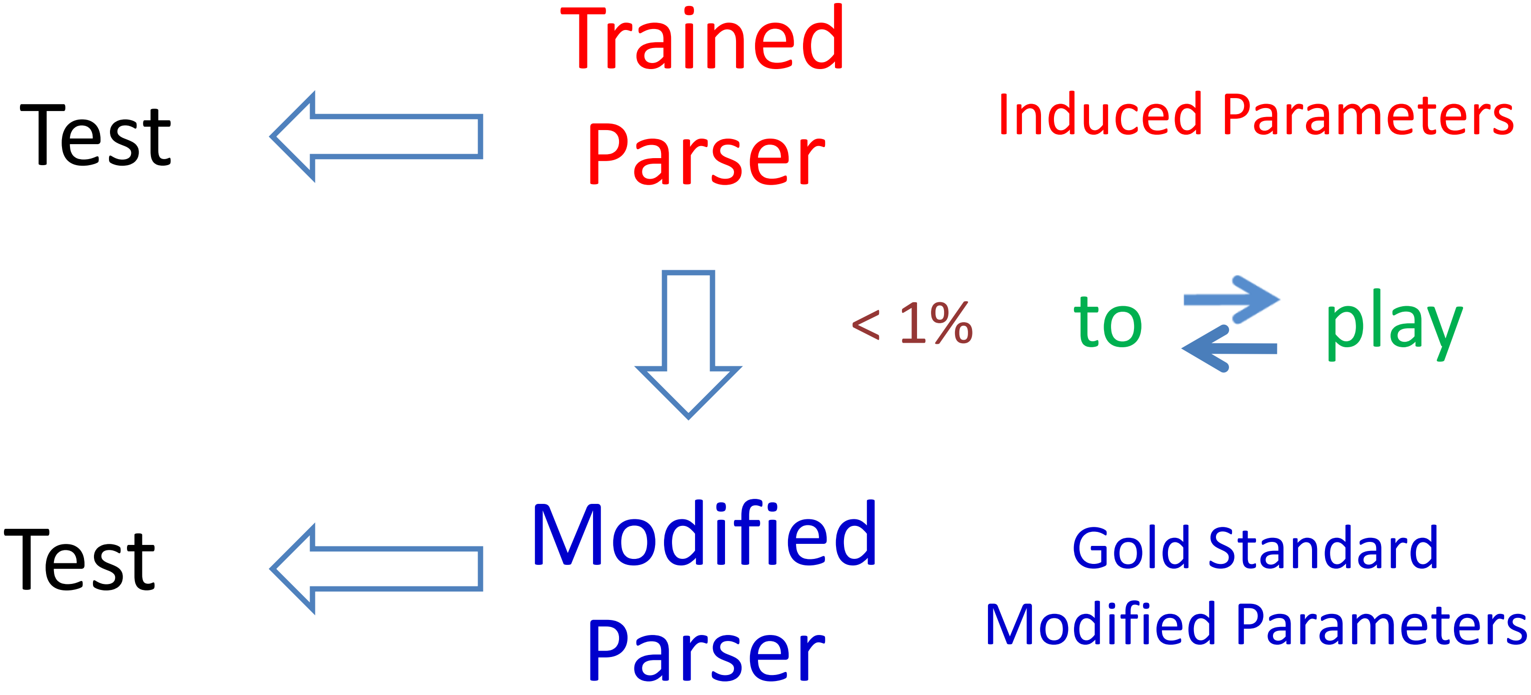


Problematic Structures

3 Different Gold Standards

Very Frequent

Sensitivity to the Annotation of Problematic Structures



Test

Trained Parser

Induced Parameters

< 1%

to ⇄ play

Test

Modified Parser

Gold Standard Modified Parameters

X 3 leading Parsers

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

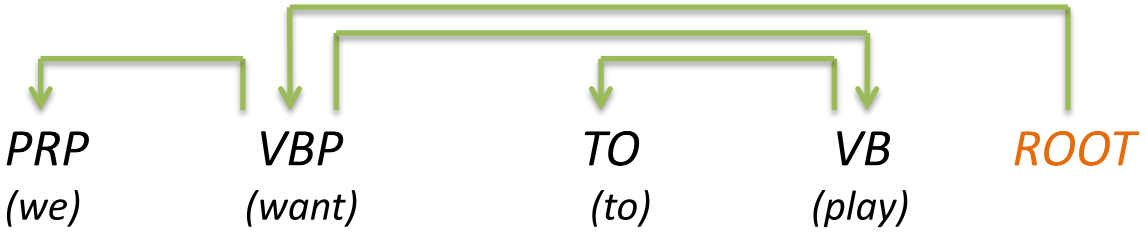
- **A possible solution** – a measure indifferent to edge directions

- How about *undirected evaluation*?

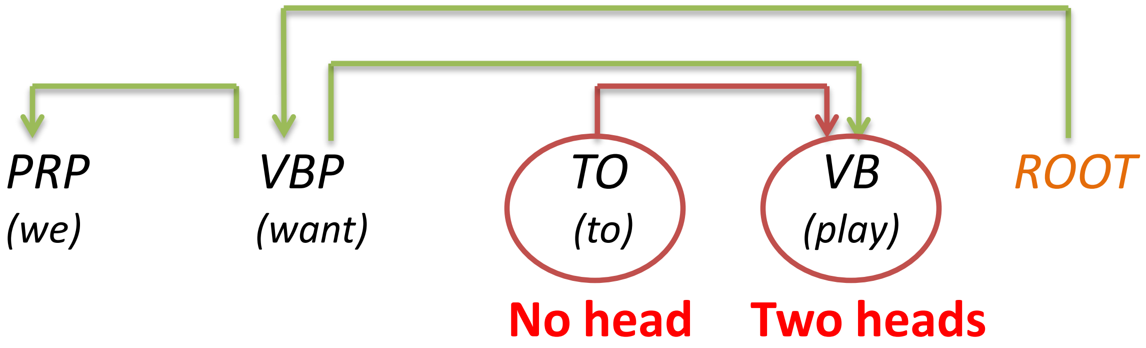
A Possible Solution

Undirected Evaluation

- Gold standard:



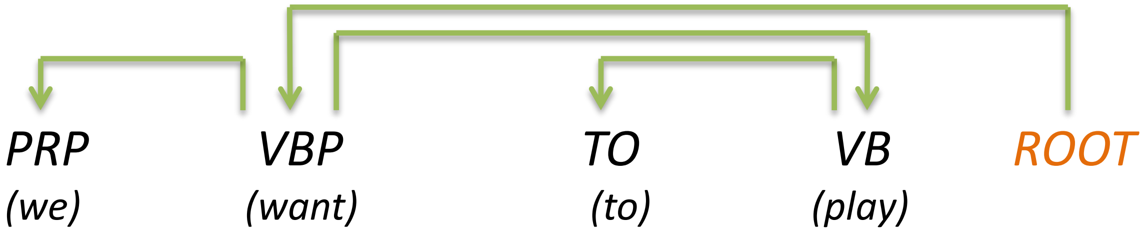
- Induced parse, with a flipped edge



A Possible Solution

Undirected Evaluation

- Gold standard:



PRP *VBP* *TO* *VB* *ROOT*

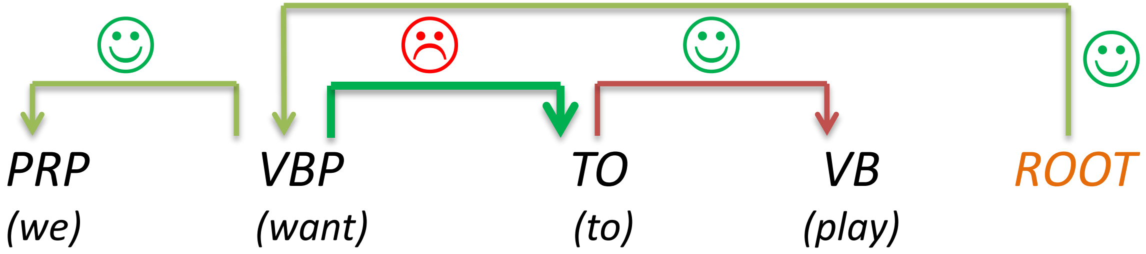
(we) *(want)* *(to)* *(play)*

- Induced parse, with a flipped edge

3/4 (75%)

undirected score

This is the minimal modification!



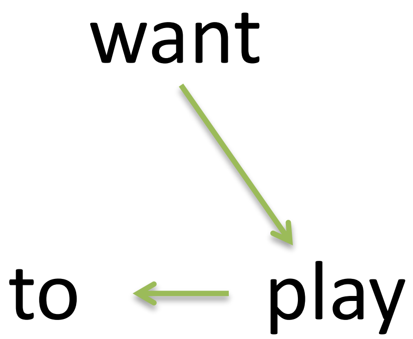
☺ ☹ ☺ ☺

PRP *VBP* *TO* *VB* *ROOT*

(we) *(want)* *(to)* *(play)*

The Neutral Edge Direction (NED) Measure

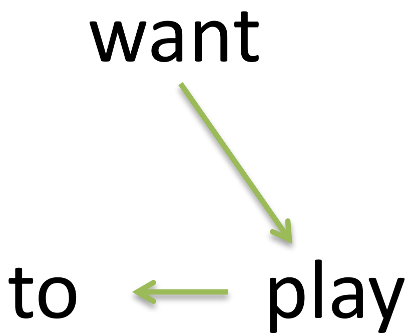
- Undirected accuracy is *not indifferent* to edge flipping
- We will now present a measure that is – *Neutral Edge Direction* (*NED*)
  - A simple extension of the undirected evaluation measure
  - Ignores edge direction flips



want

to play

Gold Standard

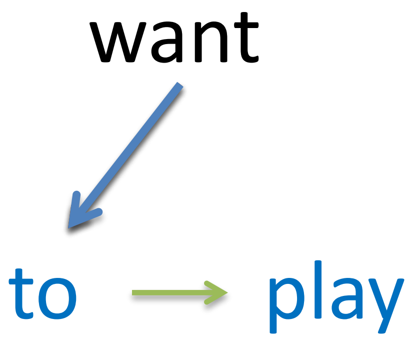


want

to play

Induced parse I
(agrees with gold std.)

- correct undirected
- correct NED attachment

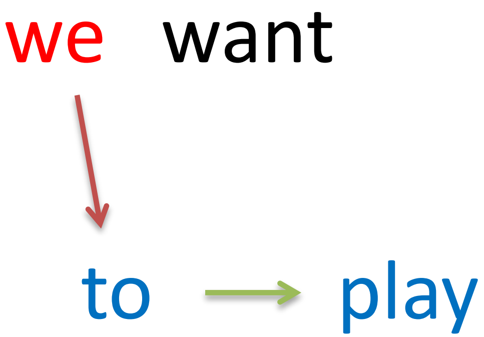


want

to play

Induced parse II
(linguistically plausible)

- undirected error
- correct NED attachment



we want

to play

Induced parse III
(linguistically implausible)

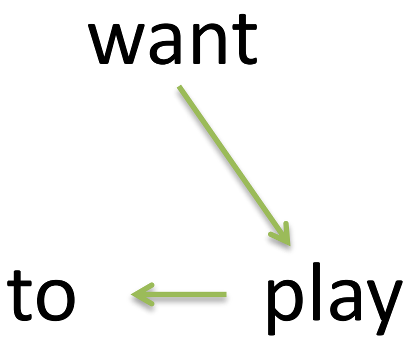
- undirected error
- NED error

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

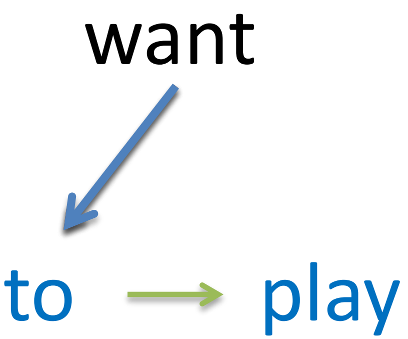
Undirected


want

to

play

Gold Standard


want

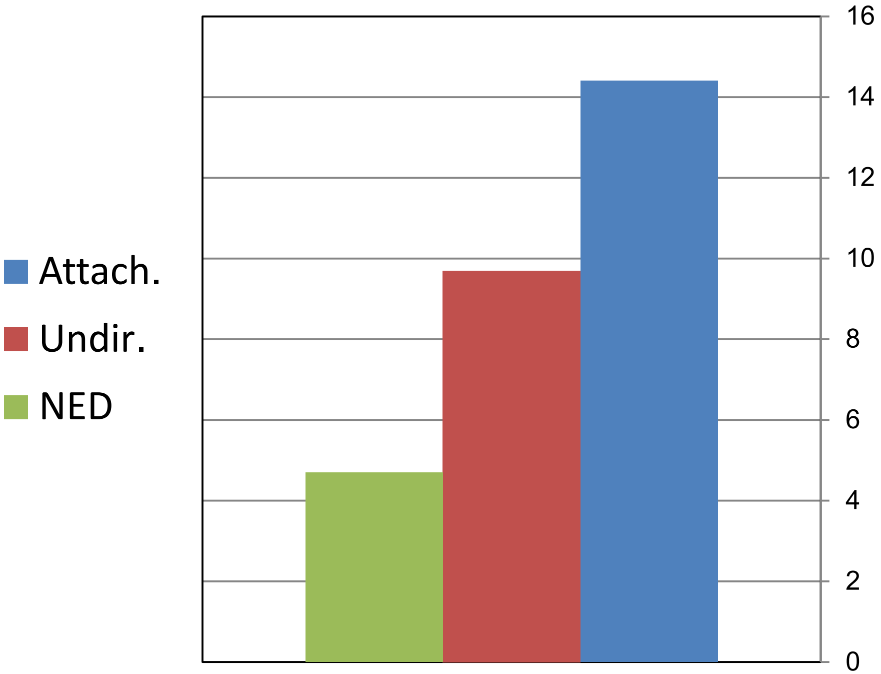
to

play

linguistically plausible parse

NED Experiments

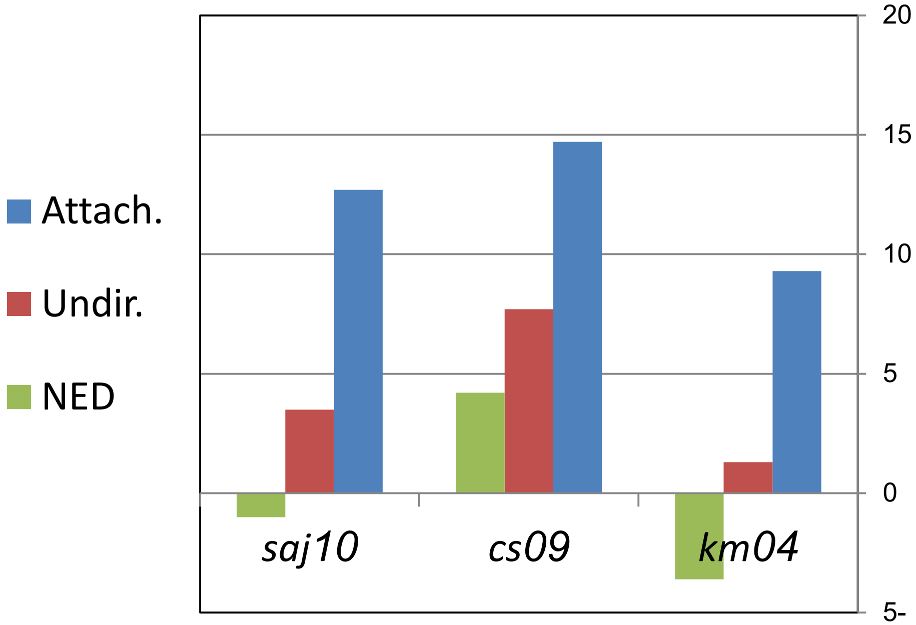
Difference Between Gold Standards



- *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 and consistent (see paper)

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%)

Take–Home Message

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

Many thanks to

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

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

NED Critiques

- *NED* is too lax
  - The edge direction *does matter* in some cases
    - E.g., “big house”: (“big” ← “house”)

- However, the standard evaluation methods are *too strict*

- *Solution*: present *both evaluation scores* in future works

NED Critiques

- *NED* only ignores structures of size 2 (e.g., “to play”)
  - What about structures of larger size (e.g., “In the house”)?

- *NED* is able to ignore some of the “wrong” size 3 annotations
  - Though not all of them

- Expanding *NED* to size 3 structures seems *too lax*

- *Possible solution*: resolve these issues in the *gold standard annotation level*

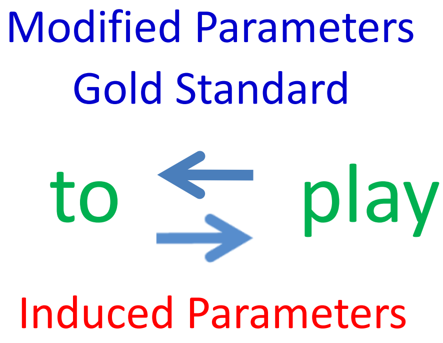
NED and Supervised Dependency Parsing

- NED is generally better suited to evaluate *unsupervised* parsers

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

Sensitivity to the Annotation of Problematic Structures

- Experimental Setup
  - 3 leading unsupervised parsers
    - All use the same parameter set
  - Training: PTB WSJ sections 2–21



Modified Parameters

Gold Standard

Induced Parameters

- Method
  - Manually modifying the learned parameters
    - Effectively *swapping edge directions* in 5 problematic structures
    - Modifications performed so to conform with the gold standard
  - Only 10–15 / ~2500 (*< 1%*) of the learned parameters are modified
  - Test (*before* and *after* modification): PTB WSJ section 23
    - Using the standard attachment score

Many thanks to

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