SEMINAR: RECENT ADVANCES IN PARSING TECHNOLOGY

Parser Evaluation Approaches
NATURE OF PARSER EVALUATION

- Return accurate syntactic structure of sentence.
  - Which representation?
- Robustness of parsing.
- Quick
- Applicable across frameworks
- Evaluation based on different sources.
  - E.g Evaluation too forgiving for same training and testing test
**Parser Evaluation**

**Intrinsic Evaluation**
- Test parser accuracy independently as “a stand-alone” system.
- Test parser output along Treebank annotations.
- BUT: High accuracy on intrinsic evaluation does not guarantee domain portability.

**Extrinsic Evaluation**
- Test accuracy of the parser by evaluating its impact on a specific NLP task. (Molla & Hunchinson 2003)
- Accuracy along frameworks and tasks.
**PARSER EVALUATION**

### Intrinsic Evaluation

- PennTreebank training & parser testing
- PARSEVAL metrics
  - PSR Bracketings
  - LA, LR,
  - LAS-UAS for dependency Parsing

### Extrinsic Evaluation

- IE Systems (PETE).
- PPI
- And more . . .
TASK-ORIENTED EVALUATION OF SYNTACTIC PARSERS & REPRESENTATIONS

Miyao, Saetre, Sagae, Matsuzaki, Tsujii (2008), Proceedings of ACL
PARSER EVALUATION ACROSS FRAMEWORKS

Parsing accuracy can’t be equally evaluated due to:
- Multiple Parsers
- Grammatical Frameworks
- Output representations: Phrase-Structure Trees, Dependency Graphs, Predicate Argument Relations.
- Training and testing along the same sources e.g.: WSJ.
**Task-oriented approach to parsing evaluation goal**

- Evaluate different syntactic parsers and their representations based on different methods.
- Measure accuracy by using an NLP task: PPI (Protein Protein Interaction).
MST
KSDEP
NO-RERANK
RERANK
BERKLEY
STANFORD
ENJU
ENJU-GENIA

PPI Extraction task

Conversion of representations

OUTPUTS
Statistical features in ML classifier
WHAT IS PPI? I

- Automatically detecting interactions between proteins.
- Extraction of relevant information from biomedical papers.
- Developed in IE Task.

Multiple techniques employed for PPI. \( \rightarrow \) effectiveness of Dependency Parsing
What is PPI? II

This study demonstrates that IL-8 recognizes and activates CXCR1, CXCR2, and the Duffy antigen by distinct mechanisms.

The molar ratio of serum retinol-binding protein (RBP) to transthyretin (TTR) is not useful to assess vitamin A status during infection in hospitalised children.

(A) <IL-8, CXCR1>

(B) <RBP, TTR>

Figure 1: CoNLL-X dependency tree
PARSERS & THEIR FRAMEWORKS

Dependency Parsing:
- MST: projective dep parsing
- KSDEP: Prob shift-reduce parsing.

Phrase Structure Parsing:
- NO-RERANK: Charniak’s (2000), lexicalized PCFG Parser.
- RERANK: Receives results from NO-RERANK & selects the most likely result.
- BERKLEY:
- STANFORD: Unlexicalized Parser
Deep Parsing
Predicate-Argument Structures reflecting semantic/syntactic relations among words, encoding deeper relations.

- ENJU: HPSG parser and extracted Grammar from Penn Treebank.
- ENJU-GENIA: Adapted to biomedical texts → GENIA
CONVERSION SCHEMES

- Convert each default parse output to other possible representations.

CoNLL: dependency tree format, easy constituent-to-dependency conversion.

PTB: PSR Trees output
  - HD: Dep Trees with syntactic heads.
  - SD: Stanford Dependency Format

PAS: Default output of ENJU & ENJU GENIA
CONVERSION SCHEMES

- 4 Representations for the PSR parsers.
- 5 Representations for the deep parsers.
DOMAIN PORTABILITY

- All versions of parsers run 2 times.
- WSJ(39832) original source
- GENIA(8127): Penn treebank style corpus of biomedical texts.

Retraining of the parsers with GENIA* to illustrate domain portability, accuracy improvements $\rightarrow$ domain adaptation
EXPERIMENTS

- Aimed corpus
- 225 biomedical paper abstracts
EVALUATION RESULTS

- Same level of achievement across WSJ trained parsers.

<table>
<thead>
<tr>
<th></th>
<th>CoNLL</th>
<th>PTB</th>
<th>HD</th>
<th>SD</th>
<th>PAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td></td>
<td>48.2/54.9/51.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>53.2/56.5/54.6</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>KSDEP</td>
<td>49.3/63.0/55.2</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NO-RERANK</td>
<td>50.7/60.9/55.2</td>
<td>45.9/60.5/52.0</td>
<td>50.6/60.9/55.1</td>
<td>49.9/58.2/53.5</td>
<td>N/A</td>
</tr>
<tr>
<td>RERANK</td>
<td><strong>53.6/59.2/56.1</strong></td>
<td>47.0/58.9/52.1</td>
<td><strong>48.1/65.8/55.4</strong></td>
<td>50.7/62.7/55.9</td>
<td>N/A</td>
</tr>
<tr>
<td>BERKELEY</td>
<td>45.8/67.6/54.5</td>
<td>50.5/57.6/53.7</td>
<td>52.3/58.8/55.1</td>
<td>48.7/62.4/54.5</td>
<td>N/A</td>
</tr>
<tr>
<td>STANFORD</td>
<td>50.4/60.6/54.9</td>
<td>50.9/56.1/53.0</td>
<td>50.7/60.7/55.1</td>
<td>51.8/58.1/54.5</td>
<td>N/A</td>
</tr>
<tr>
<td>ENJU</td>
<td>52.6/58.0/55.0</td>
<td>48.7/58.8/53.1</td>
<td>57.2/51.9/54.2</td>
<td>52.2/58.1/54.8</td>
<td>48.9/64.1/55.3</td>
</tr>
</tbody>
</table>

Table 1: Accuracy on the PPI task with WSJ-trained parsers (precision/recall/f-score)
EVALUATION RESULTS

<table>
<thead>
<tr>
<th></th>
<th>CoNLL</th>
<th>PTB</th>
<th>HD</th>
<th>SD</th>
<th>PAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>49.1/65.6/55.9</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MST</td>
<td>49.1/65.6/55.9</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>KSDEP</td>
<td>51.6/67.5/58.3</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>NO-RERANK</td>
<td>53.9/60.3/56.8</td>
<td>51.3/54.9/52.8</td>
<td>53.1/60.2/56.3</td>
<td>54.6/58.1/56.2</td>
<td>N/A</td>
</tr>
<tr>
<td>RERANK</td>
<td>52.8/61.3/56.6</td>
<td>48.3/58.0/52.6</td>
<td>52.1/60.3/55.7</td>
<td>53.0/61.1/56.7</td>
<td>N/A</td>
</tr>
<tr>
<td>BERKELEY</td>
<td>52.7/60.3/56.0</td>
<td>48.0/59.9/53.1</td>
<td>54.9/54.6/54.6</td>
<td>50.5/63.2/55.9</td>
<td>N/A</td>
</tr>
<tr>
<td>STANFORD</td>
<td>49.3/62.8/55.1</td>
<td>44.5/64.7/52.5</td>
<td>49.0/62.0/54.5</td>
<td>54.6/57.5/55.8</td>
<td>N/A</td>
</tr>
<tr>
<td>ENJU</td>
<td>54.4/59.7/56.7</td>
<td>48.3/60.6/53.6</td>
<td>56.7/55.6/56.0</td>
<td>54.4/59.3/56.6</td>
<td>52.0/63.8/57.2</td>
</tr>
<tr>
<td>ENJU-GENIA</td>
<td>56.4/57.4/56.7</td>
<td>46.5/63.9/53.7</td>
<td>53.4/60.2/56.4</td>
<td>55.2/58.3/56.5</td>
<td>57.5/59.8/58.4</td>
</tr>
</tbody>
</table>

Table 2: Accuracy on the PPI task with GENIA-retrained parsers (precision/recall/f-score)
EVALUATION RESULTS

- Dependency Parsers fastest of all.
- Deep Parsers in between speed.

<table>
<thead>
<tr>
<th></th>
<th>WSJ-trained</th>
<th>GENIA-retrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>MST</td>
<td>613</td>
<td>425</td>
</tr>
<tr>
<td>KSDEP</td>
<td>136</td>
<td>111</td>
</tr>
<tr>
<td>NO-RERANK</td>
<td>2049</td>
<td>1372</td>
</tr>
<tr>
<td>RERANK</td>
<td>2806</td>
<td>2125</td>
</tr>
<tr>
<td>BERKELEY</td>
<td>1118</td>
<td>1198</td>
</tr>
<tr>
<td>STANFORD</td>
<td>1411</td>
<td>1645</td>
</tr>
<tr>
<td>ENJU</td>
<td>1447</td>
<td>727</td>
</tr>
<tr>
<td>ENJU-GENIA</td>
<td>821</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Parsing time (sec.)
DISCUSSION
FORMALISM INDEPENDENT PARSER EVALUATION WITH CCG & DEPBank
DEPBANK

- Dependency bank, consisting of PAS Relations.
- Annotated to cover a wide selection of grammatical features.
- Produced semi-automatically as a product of XLE System

Briscoe’s & Caroll (2006) Reannotated DepBank

- Reannotation with simpler GRs.
- Original DepBank annotations kept the same.
GOAL OF THE PAPER
- Perform evaluation of CCG Parser outside of the CCG bank.
- Evaluation in DepBank.
- Conversion of CCG dependencies to Depbank GRs.
- Measuring the difficulty and effectiveness of the conversion.
- Comparison of CCG Parser against RASP Parser.
CCG PARSER

- Predicate- Argument dependencies in terms of CCG lexical categories.

- “IBM bought the company”
  $<\text{bought, (S/NP}_1\text{)}/\text{NP}_2, \text{2 company, -}>$
MAPPING OF GRs to CCG DEPENDENCIES

Measuring the difficulty transformation from one formalism to other
MAPPING OF GRs TO CCG DEPENDENCIES

2\textsuperscript{nd} Step

- Post Processing of the output by comparing CCG derivations corresponding to Depbank outputs.
- Forcing the parser to produce gold-standard derivations.
- Comparison of the GRs with the Depbank outputs and measuring Precision & Recall.

- Precision: 72.23%   Recall: 79.56%   F-score: 77.6%
- Shows the difference between schemes.
- Still a long way to the perfect conversion.
### EVALUATION WITH RASP PARSER

<table>
<thead>
<tr>
<th>Relation</th>
<th>CCG parser</th>
<th>CCGbank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Prec</td>
</tr>
<tr>
<td>aux</td>
<td>92.15</td>
<td>94.20</td>
</tr>
<tr>
<td>conj</td>
<td>72.33</td>
<td>79.73</td>
</tr>
<tr>
<td>ta</td>
<td>46.58</td>
<td>52.31</td>
</tr>
<tr>
<td>det</td>
<td>89.09</td>
<td>95.25</td>
</tr>
<tr>
<td>ncm</td>
<td>72.72</td>
<td>75.75</td>
</tr>
<tr>
<td>xmod</td>
<td>49.70</td>
<td>43.46</td>
</tr>
<tr>
<td>cmod</td>
<td>36.56</td>
<td>51.50</td>
</tr>
<tr>
<td>pmod</td>
<td>32.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ncs</td>
<td>72.61</td>
<td>83.92</td>
</tr>
<tr>
<td>xsubj</td>
<td>30.77</td>
<td>0.00</td>
</tr>
<tr>
<td>csubj</td>
<td>20.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dobj</td>
<td>81.29</td>
<td>87.03</td>
</tr>
<tr>
<td>obj2</td>
<td>26.09</td>
<td>65.00</td>
</tr>
<tr>
<td>iobj</td>
<td>73.34</td>
<td>77.60</td>
</tr>
<tr>
<td>xcomp</td>
<td>77.28</td>
<td>76.68</td>
</tr>
<tr>
<td>ccomp</td>
<td>55.55</td>
<td>79.55</td>
</tr>
<tr>
<td>pcomp</td>
<td>69.57</td>
<td>0.00</td>
</tr>
<tr>
<td>macroaverage</td>
<td>62.94</td>
<td>65.61</td>
</tr>
<tr>
<td>microaverage</td>
<td>76.29</td>
<td>82.44</td>
</tr>
</tbody>
</table>