§

Our ensemble of parsers

Research questions

iv. What accuracy do the parsers have on the training domain?

iii. What accuracy do the parsers have on the target domain?

maximum spanning trees-based parsing with annotation and subsequent cross-checking for consistency correctly?

What accuracy does our ensemble of parsers achieve on the target domain?

What accuracy do the parsers have on the training domain?

Fig. 3: 10-Fold cross-validation

Developing a gold standard on the test corpus

Two annotators annotated independently

IAA: UAS= 0.95 (±0.01), LAS= 0.93 (±0.01)

ensemble: Ens-1 = Majority vote (Match-3 /-2), default MATE except MATE = S

ensemble: Ens-2: Majority vote (Match-3), default MATE except MATE = S

ensemble: Ens-3: Majority vote (Match-3 /-2), default MATE except MATE = S

ensemble: Ens-4: Majority vote (Match-3 /-2), default MATE except MATE = S

ENS: Ens-1: Majority vote (Match-3 /-2), default MATE

ENS: Ens-2: Majority vote (Match-3), default MATE except MATE = S or OBJA, then JWCDG

Con: Ensembles do not outperform best individual parser

Pro: High reliability of ensemble majority vote

PARSER-SPECIFIC CHALLENGES

In addition: Highlight of lists and exercises

Application: Ensemble support of manual parser correction

Future work: Domain adaptation by including the gold standard in the training data

REFERENCES


The CDG Team (2001): Improving accuracy in word class tagging through the combination of machine learning systems. In: Computational Linguistics 27 (2).

Preference: Majority vote (Match-3 /-2), default MATE except MATE = S

Match-3 instances: Accuracy of parser ensemble

Score:

UAS = 0.85 (±0.02) according to Skjærholm (2014)

Fig. 1: Dependency parse CONLL format

Fig. 2: 10-fold cross-validation

Fig. 5: Accuracy evaluation

Es: Ens-1: Majority vote (Match-3 /-2), default MATE

Es: Ens-2: Majority vote (Match-3), default MATE except MATE = S or OBJA, then JWCDG

ESME-2: Majority vote (Match-3), default MATE except MATE = S

ESME-3: Majority vote (Match-3 /-2), default MATE except MATE = S

ESME-4: Majority vote (Match-3 /-2), default MATE except MATE = S

ESME: Majority vote (Match-3 /-2), default MATE

Fig. 3: Example dependency parse (HDT Annotviewer)

The "Multistrategy Approach"

A method of tagging texts by combining the output of a cluster of taggers

Tagger models result from training different learning algorithms on the same data

Different taggers create their analyses in different ways such that their errors are uncorrelated

Hypothesis: A reasonable weighted combination of the tagger choices can obtain better results than the individual taggers do (van Halteren et al. 2001, p. 201)

Related work: Segaard (2010), Rehbein et al. (2014)

Research questions

i. What accuracy do the parsers have on the training domain?

ii. What accuracy do the parsers have on the target domain?

iii. What accuracy does our ensemble of parsers achieve on the target domain?

iv. What kind of sentences does the ensemble fail to parse correctly?

v. To what extend can the ensemble parser support manual annotation?

The training corpus

Hamburg Dependency Treebank (HDT): Part A

Contains 101,999 sentences produced by manual annotation and subsequent cross-checking for consistency with DECCA (Dickinson & Meurers 2003; Foth et al. 2014).

Available free of charge for academic use

For more information visit: https://data.ub.uni-hamburg.de/DECCA

The test corpus

Texts from the Jena Textbook Corpus (unpublished)

Three geography textbooks: three double pages

Overall 144 sentences parsed

Our ensemble of parsers

MALT parser (Nivre et al. 2006): Transition-based parsing

MATE parser (The CDG Team 1997-2015): Second order maximum spanning trees-based parsing

JWCDG parser (The CDG Team 1997-2015): Weighted hand-written rules; developed on the basis of HDT

Overall training set: 144 sentences:

Fig. 4: Accuracy evaluation

Table 1: Scores

<table>
<thead>
<tr>
<th>ID</th>
<th>Token</th>
<th>Parser-1</th>
<th>Parser-2</th>
<th>Parser-3</th>
<th>Gold</th>
<th>Check</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Es</td>
<td>2 S</td>
<td>2 S</td>
<td>2 S</td>
<td>2 S</td>
<td>Check</td>
</tr>
<tr>
<td>2</td>
<td>ist</td>
<td>0 S</td>
<td>0 S</td>
<td>0 S</td>
<td>0 S</td>
<td>Check</td>
</tr>
<tr>
<td>3</td>
<td>ein</td>
<td>4 DET</td>
<td>4 DET</td>
<td>4 DET</td>
<td>4 DET</td>
<td>Check</td>
</tr>
<tr>
<td>4</td>
<td>Beispielz</td>
<td>2 S</td>
<td>2 PREP</td>
<td>2 PREP</td>
<td>2 PREP</td>
<td>Check</td>
</tr>
<tr>
<td>5</td>
<td>Fakt</td>
<td>0 S</td>
<td>0 S</td>
<td>0 S</td>
<td>0 S</td>
<td>Check</td>
</tr>
</tbody>
</table>

Task

Creating a syntactically annotated textbook corpus for linguistic research

This poster: Combining automatic and manual dependency annotation to reduce manual workload

Dependency Parsing

Fig. 5: Accuracy evaluation

Fig. 4: Accuracy evaluation

UAS: Unlabeled Attachment Score

LAS: Labeled Attachment Score

Fig. 3: 10-Fold cross-validation

Fig. 2: 10-Fold cross-validation

Developing a gold standard on the test corpus

Two annotators annotated independently

IAA: UAS= 0.95 (±0.01), LAS= 0.93 (±0.01)

ensemble: Ens-1 = Majority vote (Match-3 /-2), default MATE except MATE = S

ensemble: Ens-2: Majority vote (Match-3), default MATE except MATE = S or OBJA, then JWCDG

Evaluation

Scores:

UAS = 0.85 (±0.02) according to Skjærholm (2014)

Ens: Ens-1: Majority vote (Match-3 /-2), default MATE

Ens: Ens-2: Majority vote (Match-3), default MATE except MATE = S or OBJA, then JWCDG

Scores:

UAS = 0.85 (±0.02) according to Skjærholm (2014)

Ens: Ens-1: Majority vote (Match-3 /-2), default MATE

Ens: Ens-2: Majority vote (Match-3), default MATE except MATE = S or OBJA, then JWCDG

Fig. 4: Accuracy evaluation

Graph 1: Accuracy evaluation

Pro: Highlighting of perfect labels (AUX, AVZ, DET, GMOD, OBJA, SUBJ) and complete sentence matches

CONCLUSIONS

Pro: High reliability of ensemble majority vote

Can: Ensembles do not outperform best individual parser

Application: Ensemble support of manual parser correction

Future work: Domain adaptation by including the gold standard in the training data

REFERENCES


