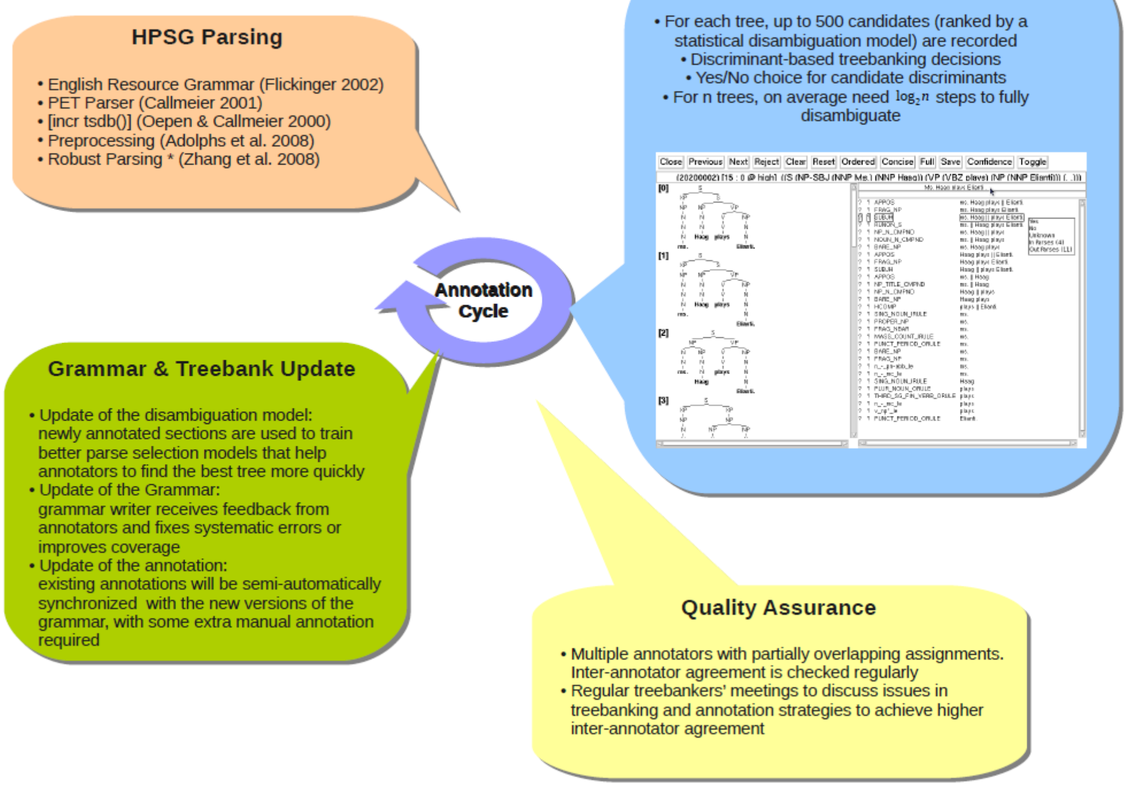
Discriminants-based Treebanking of the WSJ sections of the Penn Treebank

Valia Kordoni, Yi Zhang, Iliana Simova

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

UNIVERSITÄT DES SAARLANDES



Goal

To (re-)treebank all sentences in the WSJ sections of the PTB with the latest version of ERG. It allows for a closer comparison with other parsing studies, and contributes independent deep syntactic and semantic annotations to (probably) the most annotated set of linguistically non-trivial real texts. There are initiatives to do the treebanking in multiple frameworks, also with the same texts translated into different languages.

Setup

- Started in October 2008
- Multiple annotators (2 on average)
- Intensive and long initial training period ($\sim$ half a year)
- Annotation speed peaked at around 35 sentences per hour

Discriminants-Based Treebanking

- Discriminant-based treebanking is efficient
- Observation shows more annotation time is spent on searching for the easily decidable discriminants
- On average $\sim$100 deriviation-tree based syntactic discriminants per 500-tree forest
- Only 6% are manually disambiguated

Discriminant Ranking

- Gather annotation log (decisions either manually made, or inferred)
- Train a MaxEnt ranking model

$$P(d|Y,D) = \frac{exp(\sum_{i=1}^{k} \lambda_i f_i(d,Y))}{\sum_{d' \in D} exp(\sum_{i=1}^{k} \lambda_i f_i(d',Y))} \quad (1)$$

- Sort discriminants according to their chances of being manually decided

Features

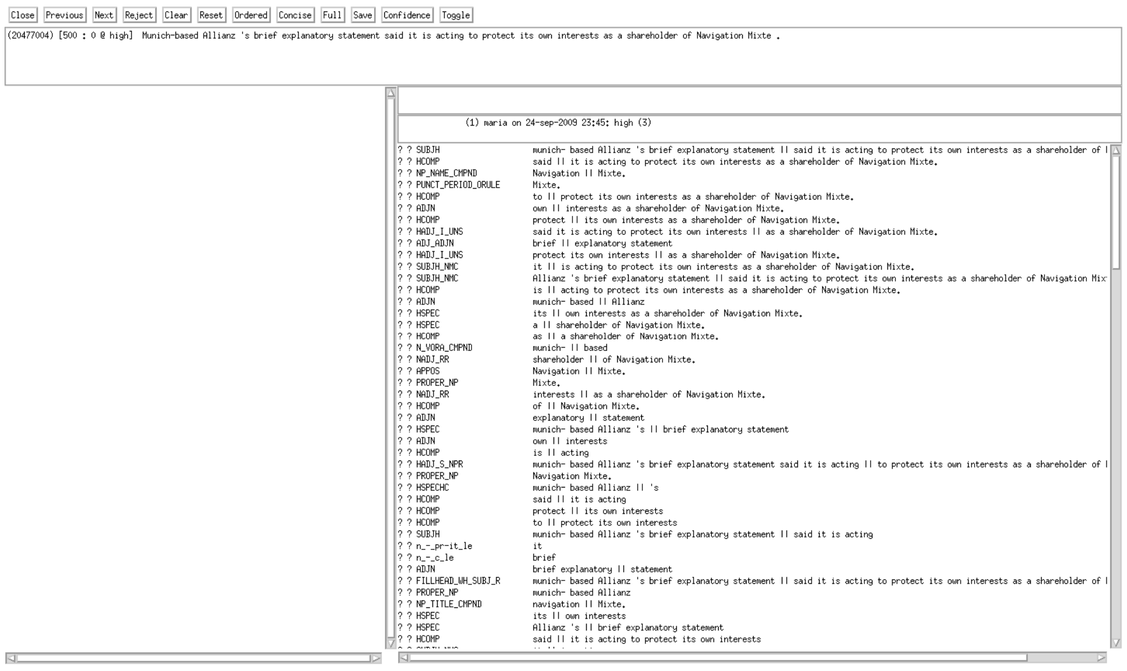
Feature	Possible Values	Example
discriminant type	RULE/LEX	RULE
edge position	FULL/FRONT/BACK	FULL
edge span	*length*(constituent)/*length*(sentence)	4/4
edge category	rule or lexical type name	SUBJH
level of discrimination	$\sum_{y \in Y} \delta(y)/\lvert Y \rvert$	4/15
branch splitting	*length*(left-dtr)/*length*(constituent)	2/4

Evaluation

- Multiple timed annotation sessions with multiple annotators
- Annotating PARC 700
- Using different discriminant ranking models
- Measure both annotation speed and inter-annotator agreement

Before ...

... and after Discriminant Ranking



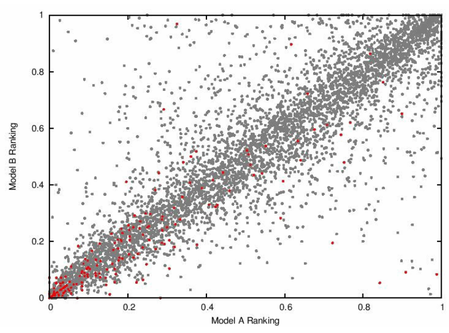
Results

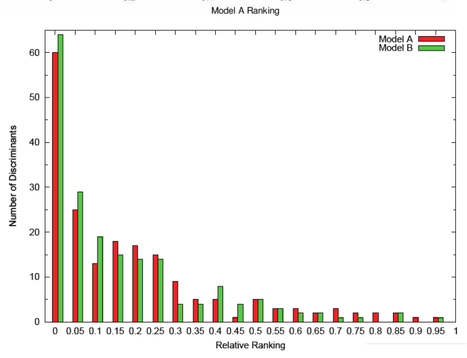
Speed

Ranking Model	Speed (s/h)	Speed-up (%)
Baseline	61.9	–
Own-model	**96.1**	**55%**
Peer-model	94.6	53%
Joint-model	95.0	53%

ITA

Ranking Model	Cohen's KAPPA (κ)
Baseline	0.5404
Own-model	**0.5798**
Peer-model	0.5567
Joint-model	0.5536





Conclusions and Outlook

- We are nearing the end: Done annotating sections 00-20
- Treebanking efficiency can be improved signicantly by discriminant ranking without hurting annotation quality
- Discriminant ranking models for training new annotators