

Project Update: WSJ Treebanking with ERG

DELPH-IN Summit 2010 Paris

Yi Zhang Valia Kordoni

German Research Center for Artificial Intelligence
Saarland University, Saarbrücken, Germany



Project Goal

To (re-)treebank all sentences in the **WSJ** sections of the PTB

Project Goal

To (re-)treebank all sentences in the **WSJ** sections of the PTB with the latest version of **ERG**.

Project Goal

To (re-)treebank all sentences in the **WSJ** sections of the PTB with the latest version of **ERG**. It will allow for a closer **comparison** with other parsing studies,

Project Goal

To (re-)treebank all sentences in the **WSJ** sections of the PTB with the latest version of **ERG**. It will allow for a closer **comparison** with other parsing studies, and contribute **independent** deep syntactic and semantic **annotation** to (probably) the most annotated set of linguistically non-trivial real texts.

Project Goal

To (re-)treebank all sentences in the **WSJ** sections of the PTB with the latest version of **ERG**. It will allow for a closer **comparison** with other parsing studies, and contribute **independent** deep syntactic and semantic **annotation** 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

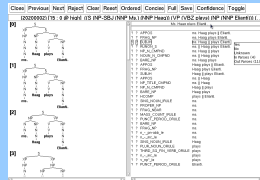
Annotation Iterations

HPSG Parsing

- English Resource Grammar (Flickinger 2002)
- PET Parser (Callmeier 2001)
- [incr tsdb()] (Oepen & Callmeier 2000)
- Preprocessing (Adolphs et al. 2008)
- Robust Parsing * (Zhang et al. 2008)

Treebanking

- For each tree, up to 500 candidates (ranked by a statistical disambiguation model) are recorded
 - Discriminant-based treebanking decisions
 - Yes/No choice for candidate discriminants
- For n trees, on average need $\log_{2,n}$ steps to fully disambiguate



Annotation Cycle

Grammar & Treebank Update

- Update of the disambiguation model:
newly annotated sections are used to train better parse selection models that help annotators to find the best tree more quickly
- Update of the Grammar:
grammar writer receives feedback from annotators and fixes systematic errors or improves coverage
- Update of the annotation:
existing annotations will be semi-automatically synchronized with the new versions of the grammar, with some extra manual annotation required

Quality Assurance

- Multiple annotators with partially overlapping assignments. Inter-annotator agreement is checked regularly
- Regular treebankers' meetings to discuss issues in treebanking and annotation strategies to achieve higher inter-annotator agreement

Current Status

- ▶ Done annotating sections 00-12 ($\sim 50\%$)
- ▶ Interannotator agreement stable at $\kappa = 0.63$ (calculated on derivation trees with ParsEval)
- ▶ Currently going through the second major grammar update (0902 $\rightarrow$ 1007)
- ▶ Changing annotators

Discriminants-Based Treebanking

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

A Screenshot Saves A Thousanda Words

Close Previous Next Reject Clear Reset Ordered Concise Full Save Confidence Toggle

(20477004) [500 x 0 high] Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste .

?? FILLHEAD_UH.SUBJ_R Munich-based Allianz 's brief explanatory statement II said it is acting to protect its own interests as a shareholder of I
?? APPOS Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Nav
?? NABJ_RC_PR Munich- II based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of I
?? PROPER_MP Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Nav
?? FRAG_MP Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Nav
?? FRAG_R_ND_UH_I Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests II as a shareholder of I
?? FILLHEAD_ND_UH_I Munich-based II Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of I
?? SUBJH Munich-based Allianz 's brief explanatory statement II said it is acting to protect its own interests as a shareholder of I
?? NABJ_S_MP Munich-based Allianz 's brief explanatory statement said it is acting II to protect its own interests as a shareholder of I
?? NABJ_RC_PR Munich- II based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Nav
?? PROPER_MP Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Nav
?? HCDMP based Allianz 's brief explanatory statement said it is acting to protect its own interests II as a shareholder of Navigati
?? NABJ_I_UNE based II Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigati
?? FILLHEAD_REL Munich- II Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigati
?? FILLHEAD_REL based II Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigati
?? HSPEC Allianz 's II brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste
?? NABJ_S_MP Allianz 's brief explanatory statement said it is acting II to protect its own interests as a shareholder of Navigation Miste
?? SUBJH_NIC Allianz 's brief explanatory statement II said it is acting to protect its own interests as a shareholder of Navigation Miste
?? SUBJH_NIC Allianz 's brief explanatory statement II said it is acting to protect its own interests as a shareholder of Navigation Miste
?? NABJ_S_MP Allianz 's brief explanatory statement said it is acting II to protect its own interests as a shareholder of Navigation
?? NABJ_RC brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste,
?? NABJ_S_MP brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste,
?? DMPER brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste,
?? SUBJH_NIC explanatory statement II said it is acting to protect its own interests as a shareholder of Navigation Miste,
?? NABJ_S_MP explanatory statement said it is acting II to protect its own interests as a shareholder of Navigation Miste,
?? FILLHEAD_REL explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste,
?? NABJ_RC_PR Munich- II based Allianz 's brief explanatory statement said it is acting to protect its own interests
?? NABJ_RC_PR Munich- II based Allianz 's brief explanatory statement said it is acting to protect its own interests
?? PROPER_MP Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests
?? HCDMP based II Allianz 's brief explanatory statement said it is acting to protect its own interests
?? FILLHEAD_REL based II Allianz 's brief explanatory statement said it is acting to protect its own interests
?? EXTRA.SUBJ.F said it is acting to protect its own interests as a shareholder of Navigation Miste,
?? NABJ_I_UNE said it is acting II to protect its own interests as a shareholder of Navigation Miste,
?? NABJ_I_UNE said it is acting to protect its own interests II as a shareholder of Navigation Miste,
?? HCDMP said II it is acting to protect its own interests as a shareholder of Navigation Miste,
?? SUBJH_NIC Allianz 's brief explanatory statement II said it is acting to protect its own interests
?? HSPEC Allianz 's II brief explanatory statement said it is acting to protect its own interests
?? NABJ_S_MP Allianz 's brief explanatory statement said it is acting II to protect its own interests
?? HCDMP said II it is acting to protect its own interests as a shareholder of Navigation
?? SUBJH_NIC it II is acting to protect its own interests as a shareholder of Navigation Miste,
?? NABJ_S_MP it is acting II to protect its own interests as a shareholder of Navigation Miste,
?? SUBJH_NIC it II is acting to protect its own interests as a shareholder of Navigation
?? SUBJH_NIC it is acting to protect its own interests II as a shareholder of Navigation Miste,
?? NABJ_I_UNE is II acting to protect its own interests as a shareholder of Navigation Miste,
?? HCDMP is II acting to protect its own interests as a shareholder of Navigation Miste,

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	$length(\text{constituent})/length(\text{sentence})$	4/4
edge category	rule or lexical type name	SUBJH
level of discrimination	$\sum_{y \in Y} \delta(y)/ Y $	4/15
branch splitting	$length(\text{left-dtr})/length(\text{constituent})$	2/4

After Discriminant Ranking ...

Close Previous Next Reject Clear Reset Ordered Concise Full Save Confidence Toggle

(20477004) [500 : 0 @ high] Munich-based Allianz 's brief explanatory statement said it is acting to protect its own interests as a shareholder of Navigation Miste .

(1) maria on 24-sep-2009 23:45; high (3)

```
?? SUBJH      munich-based Allianz 's brief explanatory statement || said it is acting to protect its own interests as a shareholder of I
?? HCOMP      said || it is acting to protect its own interests as a shareholder of Navigation Miste,
?? NP_NAME_CHPND Navigation || Miste,
?? PUNCT_PERIOD_ORULE Miste,
?? HCOMP      to || protect its own interests as a shareholder of Navigation Miste,
?? ADJN       own || interests as a shareholder of Navigation Miste,
?? HCOMP      protect || its own interests as a shareholder of Navigation Miste,
?? HADJ_JUNS   said it is acting to protect its own interests || as a shareholder of Navigation Miste,
?? ADJ_ADJN    brief || explanatory statement
?? HADJ_JUNS   protect its own interests || as a shareholder of Navigation Miste,
?? SUBJH_INCL it || is acting to protect its own interests as a shareholder of Navigation Miste,
?? SUBJH_INCL Allianz 's brief explanatory statement || said it is acting to protect its own interests as a shareholder of Navigation Miste
?? HCOMP      is || acting to protect its own interests as a shareholder of Navigation Miste,
?? ADJN       munich-based || Allianz
?? HSPEC      its || own interests as a shareholder of Navigation Miste,
?? HSPEC      a || shareholder of Navigation Miste,
?? HCOMP      as || a shareholder of Navigation Miste,
?? NP_VORL_CHPND munich- || based
?? NGJ_SR      shareholder || of Navigation Miste,
?? HPOS        Navigation || Miste,
?? PROPER_MP    Miste,
?? NGJ_SR      interests || as a shareholder of Navigation Miste,
?? HCOMP      of || Navigation Miste,
?? ADJN       explanatory || statement
?? HSPEC      munich-based Allianz 's || brief explanatory statement
?? ADJN       own || interests
?? HCOMP      is || acting
?? HADJ_SAMPR  munich-based Allianz 's brief explanatory statement said it is acting || to protect its own interests as a shareholder of I
?? PROPER_MP   Navigation Miste
?? HSPCHC      munich-based Allianz || 's
?? HCOMP      said || it is acting
?? HCOMP      protect || its own interests
?? HCOMP      to || protect its own interests
?? SUBJH       munich-based Allianz 's brief explanatory statement || said it is acting
?? n__pr__tle it
?? n__c__le    brief
?? ADJN        brief explanatory || statement
?? FILLHEAD_INCLSUBJ_R munich-based Allianz 's brief explanatory statement || said it is acting to protect its own interests as a shareholder of I
?? PROPER_MP    munich-based Allianz
?? NP_TITLE_CHPND navigation || Miste,
?? HSPEC      its || own interests
?? HSPEC      Allianz 's || brief explanatory statement
?? HCOMP      said || it is acting to protect its own interests
```

Evaluation

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

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%

Results

Speed

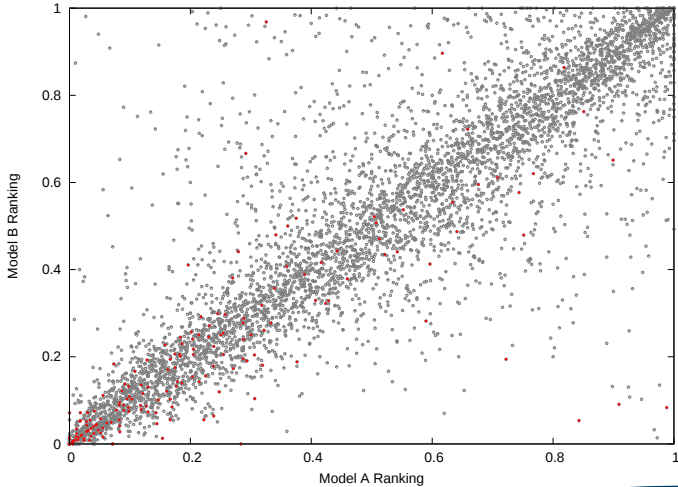
Ranking Model	Speed (s/h)	Speed-up (%)
Baseline	61.9	–
Own-model	96.1	55%
Peer-model	94.6	53%
Joint-model	95.0	53%

ITA

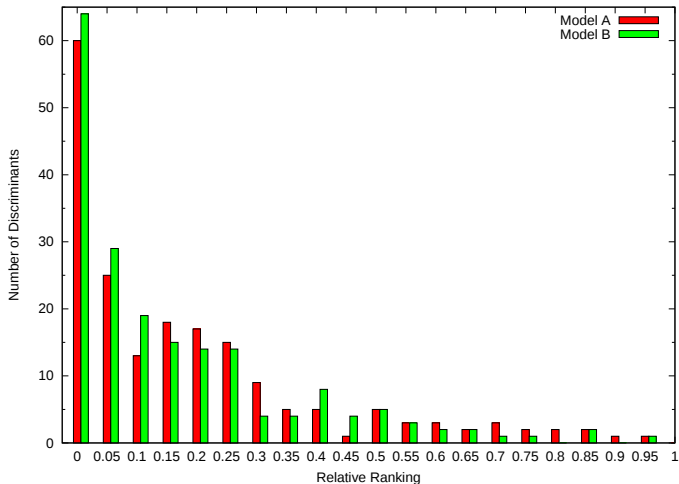
Ranking Model	Cohen's KAPPA (κ)
Baseline	0.5404
Own-model	0.5798
Peer-model	0.5567
Joint-model	0.5536



Ranking with Different Models



Ranking with Different Models



Availability

- ▶ <http://www.coli.uni-saarland.de/~yzhang/files/d-rank.tgz>

```
TSNLP(1): (load "lisp/train-discriminants.lsp")
```

```
T
```

```
TSNLP(2): (train-d '${PROFILES}) "${MODELFILE}")
```

```
TSNLP(4): (load "lisp/sort-discriminants.lsp")
```



Conclusion & Future Work

- ▶ We are making progress . . .
- ▶ Treebanking efficiency can be improved significantly by discriminant ranking without hurting annotation quality
- ▶ Discriminant ranking models for training new annotators