

# Evaluating a hybrid parser on Penn Treebank dependency triplets

Eva D'hondt, Nelleke Oostdijk, Suzan Verberne, Lou Boves  
Centre for Language and Speech Technology/ Information Foraging Lab  
Radboud University Nijmegen,  
The Netherlands  
e.dhondt@let.ru.nl

## Introduction and background

*Aim:* First test (unsupervised) of the hybrid version of the AEGIR parser.

In the **Text Mining for Intellectual Property (TM4IP)** project we aim to generate linguistic resources for data mining in the context of patent retrieval.

### The AEGIR parser:

One of these resources is the **AEGIR hybrid dependency parser** which converts texts into **Dependency Triplets (DTs)**, e.g.

```
[President_Obama, SUBJ, defend], [defend, OBJ, reform],  
[reform, ATTR, healthcare]
```

The parser uses three information sources:

1. Hand-written grammar rules (with penalties);
2. A lexicon of terms from general English and technical texts + their frequency in a corpus;
3. The DT database which contains DTs that are representative for the domain the parser is applied to + their frequency in the corpus.

### Hybridization of the parser:

Information about the **frequency** of lexicon terms or DTs guides the parsing process:

For example,

'Time flies like an arrow.'

has two different interpretations:

```
[ NP time flies ] [ VP [ V like ] [ NP an arrow ] ] or  
[ NP time ] [ VP [ V flies ] [ PP like [ NP an arrow ] ] ]
```

Toy example of DT frequency information:

```
52 [time,SUBJ,fly]      31 [fly,SUBJ,like]  
12 [fly,PREPlike,arrow] 1 [like,OBJ,arrow]
```

Toy example of frequency information in the lexicon:

```
"flies" V("fly",sing,third) 10  
"flies" N("fly",plur)      10
```

Parser calculates influence of DTs frequencies on parser rule penalties using a **semi-logarithmic scale**.

## Data

- **Evaluation Standard:** hand-made set of correct DTs of 60 sentences from the Wall Street Journal (WSJ) collection in the Penn Treebank (762 DTs);
- **DT database** for experiment 1:  
We converted the Evaluation Standard DTs to a DT database;
- **DT database** for experiment 2:
  - Using python scripts and an agfl grammar, we extracted 75,000 DTs from the syntactically annotated WSJ (minus the 60 Evaluation sentences). We obtained DTs for about 5% of the Penn Treebank.
  - Relations extracted: OBJ, ATTR, PRED, PREP, MOD, QUANT;
  - Problems during extraction:
    - Differences in description model, e.g. different places of attachment, scope, active/passive conversion, ...
    - Use of empty elements in Penn annotations

## Experimental Set-up

### Experiment 1:

*Aim:* Verification of **operational correctness** of the hybrid parser.

Comparison between:

*Baseline:* Rule-based version of the parser, no DTs in the DT database.

*Test:* DT database contains all DTs from the Evaluation Standard, i.e. train data = test data.

### Experiment 2:

*Aim:* First test of parser performance on **unseen text from same domain**.

Comparison between:

*Baseline:* Same as in experiment 1

*Test:* DT database contains 75,000 DTs from the entire Penn treebank (minus the 60 test sentences)

### Evaluation:

Parse the 60 test sentences and compare output (DTs) against DTs in Evaluation Standard for these sentences;  
If output DT = Evaluation Standard DT, then it is judged to be correct.

## Results

| Measure    | Baseline | Test (Exp 1) | Test (Exp 2) |
|------------|----------|--------------|--------------|
| Precision  | 0.50     | 0.61         | 0.54         |
| Recall     | 0.57     | 0.69         | 0.61         |
| F1-measure | 0.53     | 0.65         | 0.58         |

## Conclusions and Future Work

### Experiment 1:

- Providing the parser with the correct frequency information had a **positive influence** on parser performance.
- Why no 100% ?
  - Gaps in grammatical rules?
  - Influence of DTs < penalties in parser rules when DT has low frequency.
- Future work:
  - **Extend** parser's coverage.

### Experiment 2:

- Providing the parser with domain frequency information did not lead to a large improvement.
- Why?
  - Data sparseness; out of 75,000 DTs only a few DTs occur (potentially) in the test sentences; Much **larger DT database** needed.
  - Only information on potential DTs but **no penalizing** of incorrect DTs.
- Future work:
  - **Extend DT database** for general English domain by extracting DTs from other corpora.
  - Manually produce **blacklist** (= DTs that are not allowed), based on frequent parsing errors.