We annotate multiword expressions (MWEs) in Prague Dependency Treebank (1). We use PML scheme to capture the annotation of whole PDT as well as of our MWEs (2). Our annotators work with an annotation tool (3) and mark occurrences in sentences, which are generated from the deep syntactic layer (t-layer) of PDT: see 4, 5, 6.

Our aim is to improve a representation of MWEs (incl. named entities) in t-trees (3), because current approach is considered to be insufficient (3).

Can word sense disambiguation help statistical machine translation?

load + automatic pre-annotation (from SemLex)

one-key shortcuts for annotation

searchable SemLex browser and editor

Inter-annotator Agreement

Each t-node may be: (i) annotated with a SemLex entry (ii) one of nine types of named entities or (iii) not annotated. This yields a scale between full agreement and none. Each type of agreement is assigned a weight according to approximate amount of information it provides (4).

Then we use the slightly modified pi measure (5) on these weighted values to compute inter-annotator agreement.

Agreement measure

\[ \pi_w = \frac{A_w - A_e}{\hat{U} - A_e} \]

\[ \hat{U} = \frac{n_{A,B}}{N} + 0.052 \cdot \frac{N - n_{A,B}}{N} = 0.215 \]

\[ \pi_w = \frac{A_w - A_e}{\hat{U} - A_e} = 0.160 - 0.047 \]

\[ 0.215 - 0.047 = 0.676 \]

Annotated t-nodes

<table>
<thead>
<tr>
<th>Class c</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of t-nodes (n)</td>
<td>10,527</td>
<td>2,365</td>
<td>329</td>
<td>83,287</td>
<td>3,988</td>
</tr>
<tr>
<td>Weight (w)</td>
<td>1</td>
<td>0.5</td>
<td>0.25</td>
<td>0.052</td>
<td>0</td>
</tr>
</tbody>
</table>

Current State

- We use tectogrammatical tree structures of MWEs for automatic pre-annotation.
- The richer the tectogrammatical annotation the better the possibilities for automatic pre-annotation, which minimises human errors.
- Weighted measure that accounts for partial agreement as well as estimation of maximal agreement.
- The resulting \(\pi_w = 0.676\) is statistically significant.
- Agreement should gradually improve as:
  - we clean up the annotation lexicon
  - more entries are pre-annotated automatically
  - and further types of pre-annotation are employed.