Syntactic Identification of Occurrences of MWE in Text using a Lexicon with Dependency Structures

Eduard Bejček, Pavel Straňák, Pavel Pecina

Objective
- We have SemLex – lexicon of all MWEs in PDT 2.5
  - basic (quotation) forms
  - lemmatised forms
  - dependency structures

- How to find MWEs from SemLex in new texts?

Datasets
- Prague Dependency Treebank 2.5
  - full manual annotation
    - morphology (m), surface syntax (a), deep syntax (t)
    - MWE
  - automatic analysis: (m), (a), (t)
  - Czech National Corpus
    - SYN2006-PUB – automatic

Prague Dependency Treebank 2.5

Annotation
- no nesting of MWEs
- criteria:
  - mainly non-compositionality
  - other: translation, variability

- hypothesis 1:
  - the same MWE → the same tree structure

Automatic analysis
- it can find only MWEs in SemLex
- no semantics
- no literal meaning of a figurative MWE in SemLex

- hypothesis 2:
  - the same tree structure → the same MWE

Experiments on PDT and CNC

<table>
<thead>
<tr>
<th>layer</th>
<th>PDT-man</th>
<th>PDT-auto</th>
<th>CNC-auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>tectogrammatical</td>
<td>62 / 96</td>
<td>63 / 86</td>
<td>44 / 58</td>
</tr>
<tr>
<td>analytical</td>
<td>66 / 89</td>
<td>66 / 82</td>
<td>45 / 60</td>
</tr>
<tr>
<td>morpho, 2 words window</td>
<td>68 / 80</td>
<td>68 / 79</td>
<td>52 / 56</td>
</tr>
<tr>
<td>morpho, 3 words window</td>
<td>63 / 91</td>
<td>63 / 90</td>
<td>47 / 60</td>
</tr>
<tr>
<td>morpho, unlimited window</td>
<td>35 / 93</td>
<td>35 / 93</td>
<td>35 / 61</td>
</tr>
<tr>
<td>P / R</td>
<td>P / R</td>
<td>P / R</td>
<td>P / R</td>
</tr>
</tbody>
</table>

MWE annotation

P/R

Discussion and Conclusions
- tree structures in SemLex are too simple (→ add prepositions etc.)
- more general tectogrammatical lemmatisation should help
- 50,000 sentences in PDT data but only 546 sentences in CNC data
- there are some deficiencies in the current tectogrammatical parser
- the approach on the tectogrammatical layer is not better than on other layers, yet