Detecting Multiword Expressions by Dependency Parsing

István Nagy T. and Veronika Vincze
University of Szeged, Hungary
{nistvan,vincze}@inf.u-szeged.hu

Automatic detection of MWEs by dependency parsers in different languages

**ENGLISH verb-particle constructions**
- Penn Treebank has VPC annotation
- Bohnet and Stanford parsers were trained
- Evaluated the parsers on WikiSO, manually annotated for VPCs

**HUNGARIAN light verb constructions**
- LVCs were manually annotated in the Szeged corpus
- LVC-specific dependency relations
- Trained and evaluated the parser on Szeged corpus with 10 fold cross validation

**GERMAN light verb constructions**
- TIGER corpus has LVC annotation
- Bohnet parser was trained on TIGER
- Evaluated this model on JRC-Acquis manually annotated for LVCs

---

**English VPCs in the Penn Treebank**
- VPC:
  - Verb + particle: *show off*
  - Compositional or not
- The special relation of the verb and particle within a VPC is distinctively marked in the Penn Treebank (Marcus et al., 1993)
- It also has a specific syntactic label (PRT)
  - Turn the light off.
  - *(S (NP-SBJ *) (VP turn (NP the light) (PRT off)))*

**Automatic detection of English VPCs**
- WikiSO: full-coverage VPC annotated corpus where each individual occurrence of a VPC was manually annotated
- Examined how syntactic parsers can perform on WikiSO
- Applied the Stanford (Klein and Manning, 2003) and Bohnet (Bohnet 2010) parsers
- Only 52.57% and 58.16% of annotated VPCs on WikiSO had a verb-particle syntactic relation when we used the Stanford and Bohnet parsers
- The parsers achieved high precision scores of about 90%

<table>
<thead>
<tr>
<th>Edge type</th>
<th>Stanford</th>
<th>Bohnet</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>%</td>
<td>#</td>
</tr>
<tr>
<td>Prt</td>
<td>235</td>
<td>52.57</td>
</tr>
<tr>
<td>Prep</td>
<td>23</td>
<td>5.15</td>
</tr>
<tr>
<td>Advmod</td>
<td>56</td>
<td>12.52</td>
</tr>
<tr>
<td>Sum</td>
<td>314</td>
<td>70.24</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>1.79</td>
</tr>
<tr>
<td>None</td>
<td>125</td>
<td>27.97</td>
</tr>
<tr>
<td>sum</td>
<td>447</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**Automatic detection of Hungarian LVCs**
- The Bohnet parser was trained on the legal subdomain of the corpus.
- 10-fold cross validation was applied:
  - 86.60 (precision), 67.12 (recall), 75.63 (F-score)
- Classification: two-stage procedure (Nagy et al. 2013)
  - Extract potential LVCs
  - Classify them
- Main advantages:
  - High precision
  - Proper treatment of the non-contiguous LVCs

**German LVCs in TIGER Corpus**
- In the TIGER corpus (Brants et al. 2004), LVCs that consist of a verb and a prepositional phrase are annotated
- The PP is marked with the relation CVC (collocational verb construction)
- Verb-object pairs are excluded from the annotation
- Ablschiend nehmen “to take leave” – not an LVC here
- zur Diskussion bringen “to discuss” (zur Diskussion)LCV bringen

**Automatic detection of German LVCs**
- The Bohnet parser was trained on the TIGER corpus
- Evaluated the model on the German part of the JRC-Acquis corpus, annotated for LVCs (Rácz et al. 2014)
- 84.81 (precision), 60.91 (recall) and 70.90 (F-score)
- Same results as the English VPCs

**References**
Brants, Sabine; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; König, Esther; Leuzis, Wolfgang; Rohrer, Christian; Smith, George; Uzkoreit, Hans 2004: TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation 2, 587–620.