Inducing Discourse Marker Inventories from Lexical Knowledge Graphs

Christian Chiarcos
chiarcos@cs.uni-frankfurt.de

Applied Computational Linguistics (ACoLi), Goethe-Universität Frankfurt, Germany
Institut für Digital Humanities (IDH), Universität zu Köln, Germany

Results (also see paper)

- Direct induction (e.g., from/to English) yields best results, but depends on dictionary quality
- Apertium: "gold standard", but only 2 inventory languages with links to English
- FreeDict/MUSE: insufficient coverage
- Constrained indirect induction is a feasible fallback-strategy

- Precision is dissatisfying, but recall is reasonable => Baseline
  => Generated inventories can be a basis for manual pruning
  (note that discourse marker inventories are generally small, < 1000 entries)

Discourse marker inventories: Discourse markers => discourse relations (e.g., for Penn Discourse Treebank)

<table>
<thead>
<tr>
<th>FORM</th>
<th>level 0</th>
<th>level 1</th>
<th>level 2</th>
<th>level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>well</em></td>
<td>True</td>
<td>CONTINGENCY</td>
<td>CONTINGENCY:Cause</td>
<td>CONTINGENCY:Reason</td>
</tr>
<tr>
<td><em>weiterhin</em></td>
<td>True</td>
<td>EXPANSION</td>
<td>EXPANSION:Conjunction</td>
<td>EXPANSION:Conjunction</td>
</tr>
</tbody>
</table>

Lexical induction: Given discourse marker inventories in multiple source languages
Translate discourse markers into target languages, keep (propagate) relations
Aggregate over multiple sources => confidence scores => filters
Iterate with indirect translations

Pre-requisites: Discourse marker inventories in machine-readable (interoperable) formats
Normalization of discourse relations against a uniform relation taxonomy
Large collection of dictionaries in machine-readable (interoperable) formats

Machine-readable discourse marker inventories
http://github.com/acoli-repo/rdf4discourse/discourse-markers

- 16 languages, 19 inventories
  (Chiarcos & Ionov 2021)
- Partially building on TextLink/Connective-lex.org
- Data model: OntoLex
- Formats: RDF (=> TSV, with SPARQL)

Discourse relation taxonomies: PDTB, RST, CCR
multiple taxonomies linked with an overarching ontology
=> "translate" between frameworks, using the shortest path

Machine-readable dictionaries
http://github.com/acoli-repo/acoli-dicts

- 430+ languages, 3000+ bi-dictionaries
  (Chiarcos et al. 2020)
- RDF layer over PanLex, Apertium, FreeDict, MUSE, etc.
- Data model: OntoLex
- Formats: RDF (=> TSV, with SPARQL)

Selected subsets
- Apertium 53 dictionaries for MT, mostly Romance
- FreeDict 145 dictionaries, heterogeneous
- MUSE 108 dictionaries, machine-generated

Constrained induction
- Operate over confidence scores for discourse relations
- Initialize word w with 1/(number of senses)
- Propagate relation score to word v: average over relation scores for translations (w. score)
- Constraints: (optionally) filter by min result score
  min pivots (translations)
  min pivot languages (of translations)
  max senses (top k relations, only)

Experimental Setup
http://github.com/acoli-repo/rdf4discourse/lexical-induction

- 11 inventories, 9 languages
- PDTB, each level separately
- Evaluate prec, rec, f against target inventories
- Publish 10 induced inventories
  (Bulgarian, Greek, Esperanto, Finnish, Japanese, Norwegian, Polish, Russian, Swedish and Turkish)