Almost Total Recall: Semantic Category Disambiguation Using Large Lexical Resources and Approximate String Matching

Pontus Stenetorp*, Sampo Pyysalo†‡ Sophia Ananiadou†‡ and Jun’ichi Tsujii$

*Alzawa Laboratory, University of Tokyo | †School of Computer Science, ‡National Centre for Text Mining, University of Manchester | $Microsoft Research Asia

{pontus, smp}@is.s.u-tokyo.ac.jp, sophia.ananiadou@manchester.ac.uk, jtsujii@microsoft.com

Semantic Category Disambiguation

These findings suggest that SaIK/SaIR is requisite for the full virulence of ethnic Chinese...

Figure: Demarked textual spans

These findings suggest that SaIK/SaIR is requisite for the full virulence of ethnic Chinese...

Figure: Demarked textual spans assigned semantic categories

• Semantic Category Disambiguation: assign one or multiple semantic categories to a single continuous textual span
• Integral part of Named Entity Recognition (NER)

Research Target

• How does semantic disambiguation perform in an NLP pipeline?
• Can it minimise the number of categories exposed to an annotator?

Previous Research

• Cohen et al. (2011)
  • Define categories by ontologies
  • Associate a textual span with one or multiple categories
  • Rule-based, non-probabalistic
• Stenetorp et al. (2011)
  • Standard NER features
  • Novel large-scale fast approximate string matching with 170 databases and 20,335,426 entries
  • Single category assumption
  • Machine learning, probabalistic

Approach

• Exploit the probabalistic aspects of the model
• Use the sum of the category probabilities to threshold the number of suggestions
• Forces the model to perform a recall trade-off

Evaluation Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbreviation</th>
<th>Semantic Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioNLP/NLPBA 2004 Shared Task Corpus</td>
<td>NLPBA</td>
<td>5</td>
</tr>
<tr>
<td>Gene Regulation Event Corpus</td>
<td>SGREC</td>
<td>64 (5 collapsed)</td>
</tr>
<tr>
<td>Collaborative Annotation of a Large Biomedical Corpus</td>
<td>SSC</td>
<td>4</td>
</tr>
<tr>
<td>Epigenetics and Post-Translational Modifications</td>
<td>EPI</td>
<td>17</td>
</tr>
<tr>
<td>Infectious Diseases Corpus</td>
<td>ID</td>
<td>16</td>
</tr>
<tr>
<td>Genia Event Corpus</td>
<td>GENIA</td>
<td>11</td>
</tr>
</tbody>
</table>

Table: Corpora used for evaluation

Evaluation Metrics

• Ambiguity: average number of suggested categories
• Recall: as an ambiguity trade-off

Conclusions

• Can retain high recall while greatly reducing ambiguity
• Semantic category disambiguation is ready to support other tasks

Future Work

• Support other NLP tasks such as co-reference resolution and coordination
• Extend to Noun Phrase classification
• Does the results hold even when the amount of categories goes towards the hundreds?
• Integrate into existing annotation tool(s) as speed enhancement and quality checker

Availability

Source code, lexical resources, additional results and future research is/will be available at:

http://github.com/ninjin/simsem/

Feel free to use, derive and/or complain.

Acknowledgements

This work was supported by the Swedish Royal Academy of Sciences and by Grant-in-Aid for Specially Promoted Research (MEXT, Japan). The UK National Centre for Text Mining is funded by the UK Joint Information Systems Committee (JISC).