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

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

*Aizawa 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

東京大学 THE UNIVERSITY OF TOKYO

MANCHESTER 1824 The University of Manchester

Semantic Category Disambiguation

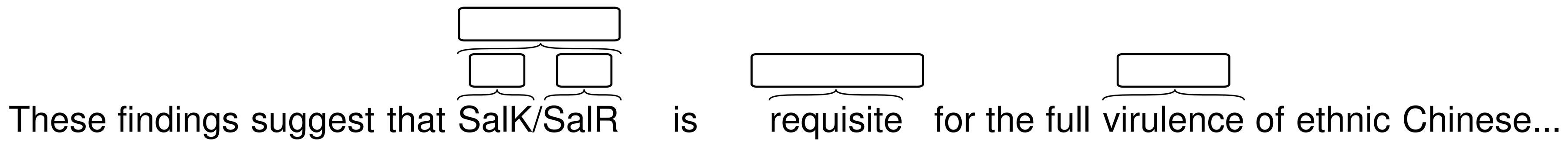

Figure: Demarked textual spans

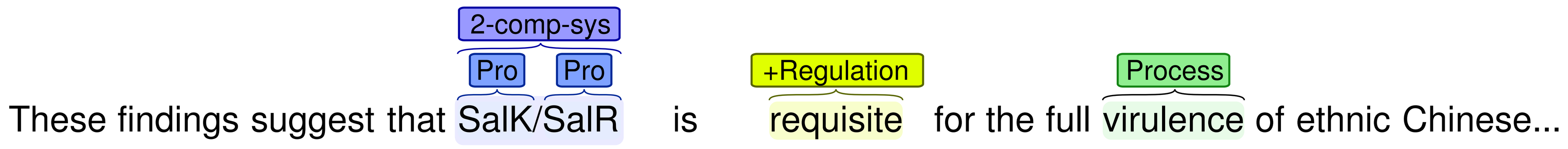

Figure: Demarked textual spans assigned semantic categories

- **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 probabalities to threshold the number of suggestions
- Forces the model to perform a recall trade-off

Experimental Results

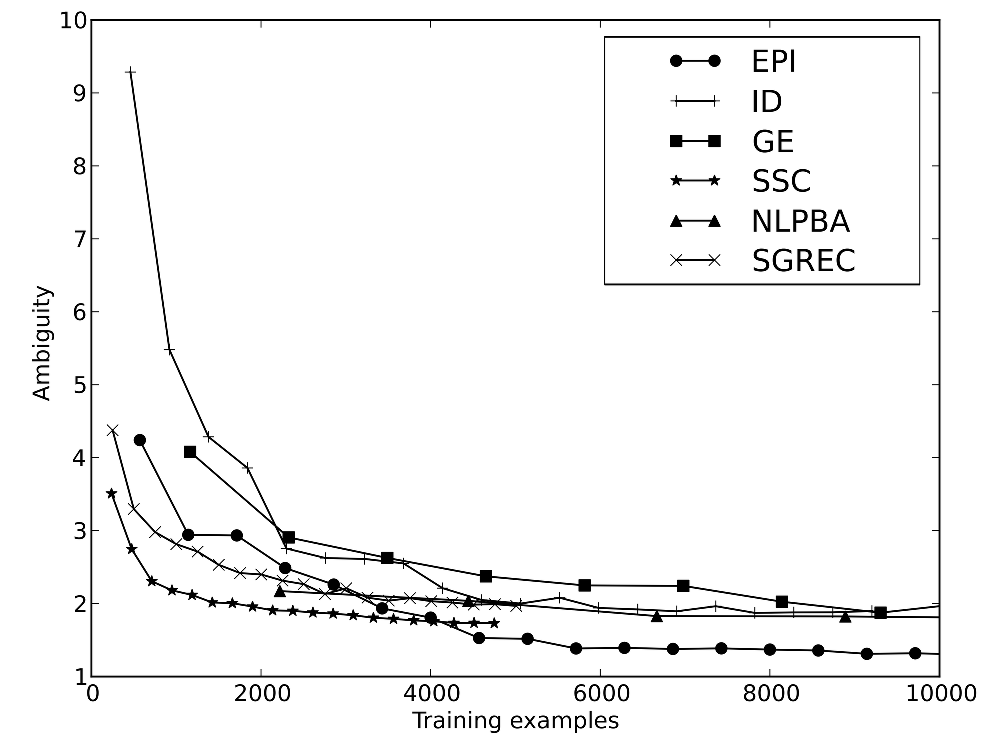

Figure: Ambiguity per dataset

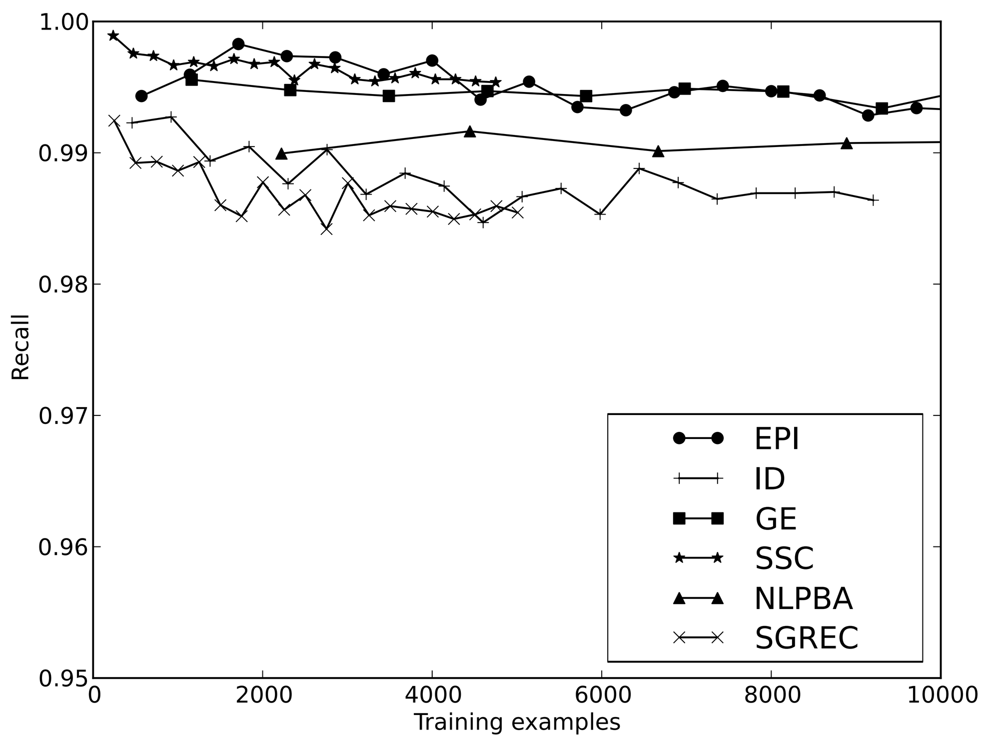

Figure: Recall per Dataset

Data set	Mean Ambiguity	Ambiguity	Mean Recall	Recall
EPI	1.8/89.4%	1.3/92.4%	99.5%	99.4%
ID	2.9/81.9%	1.9/88.1%	98.8%	98.6%
GE	2.1/80.9%	1.7/84.5%	99.4%	99.5%
SSC	2.0/50.0%	1.7/57.5%	99.6%	99.5%
NLPBA	1.8/64.0%	1.6/68.0%	99.1%	99.1%
SGREC	2.4/60.0%	2.0/66.7%	98.7%	98.6%

Table: Performance by ambiguity level/reduction and recall for the mean over the learning curve and when all training and development data was used as training data

Evaluation Datasets

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

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).