SimSem: Fast Approximate String Matching in Relation to Semantic Category Disambiguation

Pontus Stenetorp*†, Sampo Pyysalo* and Jun'ichi Tsujii‡
{*Tsujii Laboratory,†Aizawa Laboratory}, The University of Tokyo | ‡Microsoft Research Asia
{pontus,smp}@is.s.u-tokyo.ac.jp, jtsujii@microsoft.com

東京大学 THE UNIVERSITY OF TOKYO
Tsujii Laboratory

Lexical Resources for Named Entity Recognition

State-of-the-art:

- Gazetteers to boost performance (Torii et al., 2009)
- Covers few semantic categories

Problems:

- Lexical variation poses a problem for gazetteers
- No resource perfectly matches an entity class (Wang et al., 2009)

Approximate String Matching

Observations:

- We have large collections of lexical resources for various classes
- Recent advances in approximate string matching makes approximate string matching computationally feasible (Okazaki and Tsujii, 2010)

Idea:

- Approximate instead of strict gazetteers
- Use large collections to disambiguate semantic categories

SimString Features

Feature generation:

1. Query each lexical resource using cosine measure and a sliding similarity threshold [1.0, 0.7] with a step of 0.1
2. If the query returns a match, assign a feature uniquely identifying that data and threshold, and all subsequent lower thresholds

To note:

- Threshold 1.0 is equivalent to strict matching
- Cut-off motivated by the fact that low thresholds will match similarities even at a superficial level

Models

Name	Description
Internal	Span-internal features used in previous work on NER
Internal+Gazetteer	Internal features and gazetteer features
Internal+SimString	Internal features and SimString features

Table: Models used in our experimental setting

Task Setting

Given a textual span (denoted by [...]), assign the semantic category:

*[**Histone H3**]*PROTEIN *[**methylation**]*METHYLATION *at [**lys36**]*AMINOACID *was [**catalysed**]*CATALYSIS *by [**HMT**]*PROTEIN.

Experimental Results

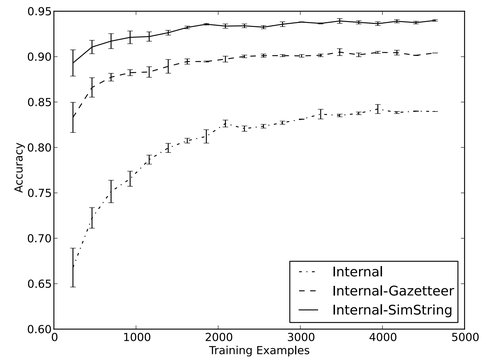

Figure: Learning curve for the CALBC CII dataset

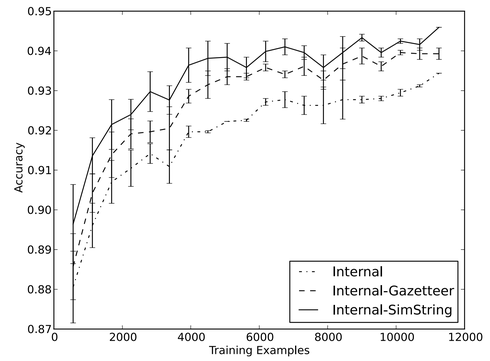

Figure: Learning curve for the EPI dataset

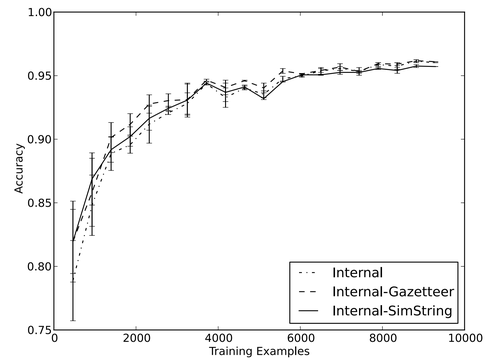

Figure: Learning curve for the ID dataset

Lexical Resources for the Experiments

Name	Semantic Categories	Entries	Lexical Resources
Gene Ontology	Multiple	128,999	12
Protein Information Resource	Proteins	691,577	1
Unified Medical Language System	Multiple	5,902,707	135
Entrez Gene	Proteins	3,602,757	5
Shi and Campagne (2005) Dictionary	Proteins	61,676	1
Jochem	Multiple	1,715,744	2
Turku Event Corpus	Multiple	4,875,964	11
Arizona Disease Corpus	Diseases	1,195	1
LINNAEUS Dictionary	Species	3,119,005	1
Webster's International Dictionary	Multiple	235,802	1
Total:	–	20,335,426	170

Table: Lexical resources gathered for our experiments

Evaluation Datasets

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

Table: Corpora used for evaluation

Conclusions

- Can not establish a clear benefit for all datasets, but works very well for one dataset
- The method appears to have potential but merits further investigations as to when it is applicable

Future Work

- Evaluate non-cosine measures for approximate string matching, which take biological knowledge into account
- Investigate as to why certain datasets although similar on a semantic category level yield different different results
- Contribution of individual resources towards overall performance

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.

BioNLP 2011 Workshop, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, June 23rd 2011