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

Pontus Stenetorp[*†] **Sampo Pyysalo**[*] **and Jun'ichi Tsujii**[‡]
[*] Tsujii Laboratory, Department of Computer Science, The University of Tokyo, Tokyo, Japan
[†] Aizawa Laboratory, Department of Computer Science, The University of Tokyo, Tokyo, Japan
[‡] Microsoft Research Asia, Beijing, People's Republic of China
{pontus,smp}@is.s.u-tokyo.ac.jp
jtsujii@microsoft.com

Abstract

In this study we investigate the merits of fast approximate string matching to address challenges relating to spelling variants and to utilise large-scale lexical resources for semantic class disambiguation. We integrate string matching results into machine learning-based disambiguation through the use of a novel set of features that represent the distance of a given textual span to the closest match in each of a collection of lexical resources. We collect lexical resources for a multitude of semantic categories from a variety of biomedical domain sources. The combined resources, containing more than twenty million lexical items, are queried using a recently proposed fast and efficient approximate string matching algorithm that allows us to query large resources without severely impacting system performance. We evaluate our results on six corpora representing a variety of disambiguation tasks. While the integration of approximate string matching features is shown to substantially improve performance on one corpus, results are modest or negative for others. We suggest possible explanations and future research directions. Our lexical resources and implementation are made freely available for research purposes at: http://github.com/ninjin/simsem

1 Introduction

The use of dictionaries for boosting performance has become commonplace for Named Entity Recognition (NER) systems (Torii et al., 2009; Ratinov and Roth, 2009). In particular, dictionaries can give an initial improvement when little or no training data is available. However, no dictionary is perfect, and all resources lack certain spelling variants and lag behind current vocabulary usage and thus are unable to cover the intended domain in full. Further, due to varying dictionary curation and corpus annotation guidelines, the definition of what constitutes a semantic category is highly unlikely to precisely match for any two specific resources (Wang et al., 2009). Ideally, for applying a lexical resource to an entity recognition or disambiguation task to serve as a definition of a semantic category there would be a precise match between the definitions of the lexical resource and target domain, but this is seldom or never the case.

Most previous work studying the use of dictionary resources in entity mention-related tasks has focused on single-class NER, in particular this is true for BioNLP where it has mainly concerned the detection of proteins. These efforts include Tsuruoka and Tsujii (2003), utilising dictionaries for protein detection by considering each dictionary entry using a novel distance measure, and Sasaki et al. (2008), applying dictionaries to restrain the contexts in which proteins appear in text. In this work, we do not consider entity mention detection, but instead focus solely on the related task of disambiguating the semantic category for a given continuous sequence of characters (a textual span), doing so we side-step the issue of boundary detection in favour of focusing on novel aspects of semantic category disambiguation. Also, we are yet to see a high-performing multi-class biomedical NER system, this motivates our desire to include multiple semantic categories.

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 136–145,
Portland, Oregon, USA, June 23-24, 2011.

2 Methods

In this section we introduce our approach and the structure of our system.

2.1 SimSem

Many large-scale language resources are available for the biomedical domain, including collections of domain-specific lexical items (Ashburner et al., 2000; Bodenreider, 2004; Rebholz-Schuhmann et al., 2010). These resources present obvious opportunities for semantic class disambiguation. However, in order to apply them efficiently, one must be able to query the resources taking into consideration both lexical variations in dictionary entries compared to real-world usage and the speed of look-ups.

We can argue that each resource offers a different view of what constitutes a particular semantic category. While these views will not fully overlap between resources even for the same semantic category, we can expect a certain degree of agreement. When learning to disambiguate between semantic categories, a machine learning algorithm could be expected to learn to identify a specific semantic category from the similarity between textual spans annotated for the category and entries in a related lexical resource. For example, if we observe the text "Carbonic anhydrase IV" marked as PROTEIN and have an entry for "Carbonic anhydrase 4" in a lexical resource, a machine learning method can learn to associate the resource with the PROTEIN category (at specific similarity thresholds) despite syntactic differences.

In this study, we aim to construct such a system and to demonstrate that it outperforms strict string matching approaches. We refer to our system as SimSem, as in "Similarity" and "Semantic".

2.2 SimString

SimString[1] is a software library utilising the CPMerge algorithm (Okazaki and Tsujii, 2010) to enable fast approximate string matching. The software makes it possible to find matches in a collection with over ten million entries using cosine similarity and a similarity threshold of 0.7 in approximately 1 millisecond with modest modern hardware. This makes it useful for querying a large collection of strings to find entries which may differ from the query string only superficially and may still be members of the same semantic category.

As an example, if we construct a SimString database using an American English wordlist[2] and query it using the cosine measure and a threshold of 0.7. For the query "reviewer" SimString would return the following eight entries: review, viewer, preview, reviewer, unreviewed, televiewer, and revieweress. We can observe that most of the retrieved entries share some semantic similarity with the query.

2.3 Machine Learning

For the machine learning component of our system we use the L2-regularised logistic regression implementation of the LIBLINEAR[3] software library (Fan et al., 2008). We do not normalise our feature vectors and optimise our models' penalty parameter using k-fold cross-validation on the training data. In order to give a fair representation of the performance of other systems, we use a rich set of features that are widely applied for NER (See Table 1).

Our novel SimString features are generated as follows. We query each SimString database using the cosine measure with a sliding similarity threshold, starting at 1.0 and ending at 0.7, lowering the threshold by 0.1 per query. If a query is matched, we generate a feature unique for that database and threshold, we also generate the same feature for each step from the current threshold to the cut-off of 0.7 (a match at e.g. 0.9 similarity also implies matches at 0.8 and 0.7).

The cut-off is motivated by the fact that very low thresholds introduces a large degree of noise. For example, for our American English wordlist the query "rejection" using threshold 0.1 and the cosine measure will return 13,455 results, among them "questionableness" which only have a single sequence "ion" in common.

It is worthwhile to note that during our preliminary experiments we failed to establish a consistent benefit from contextual features across our development sets. Thus, contextual features are not included in our feature set and instead our study focuses only

[1] http://www.chokkan.org/software/simstring/

[2] /usr/share/dict/web2 under FreeBSD 8.1-RELEASE, based on Webster's Second International dictionary from 1934

[3] We used version 1.7 of LIBLINEAR for our experiments

Feature	Type	Input	Value(s)
Text	Text	Flu	Flu
Lower-cased	Text	DNA	dna
Prefixes: sizes 3 to 5	Text	bull	bul, ...
Suffixes: sizes 3 to 5	Text	bull	ull, ...
Stem (Porter, 1993)	Text	performing	perform
Is a pair of digits	Bool	42	True
Is four digits	Bool	4711	True
Letters and digits	Bool	C4	True
Digits and hyphens	Bool	9-12	True
Digits and slashes	Bool	1/2	True
Digits and colons	Bool	3,1	True
Digits and dots	Bool	3.14	True
Upper-case and dots	Bool	M.C.	True
Initial upper-case	Bool	Pigeon	True
Only upper-case	Bool	PMID	True
Only lower-case	Bool	pure	True
Only digits	Bool	131072	True
Only non-alpha-num	Bool	#*$!	True
Contains upper-case	Bool	gAwn	True
Contains lower-case	Bool	After	True
Contains digits	Bool	B52	True
Contains non-alpha-num	Bool	B52;s	True
Date regular expression[4]	Bool	1989-01-30	True
Pattern	Text	1B-zz	0A-aa
Collapsed Pattern	Text	1B-zz	0A-a

Table 1: Basic features used for classification

the features that are generated solely from the textual span which has been annotated with a semantic category (span-internal features) and the comparison of approximate and strict string matching.

3 Resources

This section introduces and discusses the preprocessing and statistics of the lexical and corpus resources used in our experiments.

3.1 Lexical Resources

To generate a multitude of SimString databases covering a wide array of semantic categories we employ several freely available lexical resources (Table 2).

The choice of lexical resources was initially made with the aim to cover commonly annotated domain semantic categories: the CHEBI and CHEMICAL subsets of JOCHEM for chemicals, LINNAEUS for species, Entrez Gene and SHI for proteins. We then expanded the selection based on error analysis to increase our coverage of a wider array of semantic categories present in our development data.

We used the GO version from March 2011, extracting all non-obsolete terms from the ontology and separating them into the three GO subontologies: biological process (BP), cellular component (CC) and molecular function (MF). We then created an additional three resources by extracting all exact synonyms for each entry. Lastly, we expanded these six resources into twelve resources by applying the GO term variant generation technique described by Beisswanger et al. (2008).

UMLS, a collection of various resources, contain 135 semantic categories (e.g. Body Location or Region and Inorganic Chemical) which we use to create a database for each category.

For Entrez Gene we extracted all entries for the following types: gene locus, protein name, protein description, nomenclature symbol and nomenclature fullname, creating a SimString database for each. This leaves some parts of Entrez Gene unutilised, but we deemed these categories to be sufficient for our experiments.

The Turku Event Corpus is a resource created by applying an automated event extraction system on the full release of PubMed from 2009. As a precondition for the event extraction system to operate, protein name recognition is necessary; for this corpus, NER has been performed by the corpus curators using the BANNER (Leaman and Gonzalez, 2008) NER system trained on GENETAG (Tanabe et al., 2005). We created a database (PROT) containing all protein annotations, extracted all event triggers (TRIG) and created a database for each of the event types covered by the event extraction system.

For the AZDC corpus, we extracted each annotated textual span since the corpus covers only a single semantic category. Similarly, the LINNAEUS dictionary was converted into a single database since it covers the single category "species".

Table 3 contains the statistics per dictionary resource and the number of SimString databases created for each resource. Due to space requirements we leave out the full details for GO BP, GO CC, GO MF, UMLS, Entrez Gene and TURKU TRIG, and instead give the total entries for all the databases generated from these resources.

[4] A simple regular expression matching dates:
^(19|20)\d\d[- /.](0[1-9]|1[012])[- /.](0[1-9]|[12][0-9]|3[01])$
from http://www.regular-expressions.info/dates.html

Name	Abbreviation	Semantic Categories	Publication
Gene Ontology	GO	Multiple	Ashburner et al. (2000)
Protein Information Resource	PIR	Proteins	Wu et al. (2003)
Unified Medical Language System	UMLS	Multiple	Bodenreider (2004)
Entrez Gene	–	Proteins	Maglott et al. (2005)
Automatically generated dictionary	SHI	Proteins	Shi and Campagne (2005)
Jochem	JOCHEM	Multiple	Hettne et al. (2009)
Turku Event Corpus	TURKU	Proteins and biomolecular events	Björne et al. (2010)
Arizona Disease Corpus	AZDC	Diseases	Chowdhury and Lavelli (2010)
LINNAEUS Dictionary	LINNAEUS	Species	Gerner et al. (2010)
Webster's International Dictionary	WID	Multiple	–

Table 2: Lexical resources gathered for our experiments

Resource	Unique Entries	Databases
GO BP	67,411	4
GO CC	5,993	4
GO MF	55,595	4
PIR	691,577	1
UMLS	5,902,707	135
Entrez Gene	3,602,757	5
SHI	61,676	1
CHEBI	187,993	1
CHEMICAL	1,527,751	1
TURKU PROT	4,745,825	1
TURKU TRIG	130,139	10
AZDC	1,195	1
LINNAEUS	3,119,005	1
WID	235,802	1
Total:	20, 335, 426	170

Table 3: Statistics per dictionary resource

3.2 Corpora

To evaluate our approach we need a variety of corpora annotated with multiple semantic categories. For this purpose we selected the six corpora listed in Table 4.

The majority of our corpora are available in the common stand-off style format introduced for the BioNLP 2009 Shared Task (BioNLP'09 ST) (Kim et al., 2009). The remaining two, NLPBA and CALBC CII, were converted into the BioNLP'09 ST format so that we could process all resources in the same manner for our experimental set-up.

In addition to physical entity annotations, the GREC, EPI, ID and GENIA corpora incorporate event trigger annotations (e.g. Gene Regulatory Event (GRE) for GREC). These trigger expressions carry with them a specific semantic type (e.g. "interact" can carry the semantic type BINDING for GENIA), allowing us to enrich the data sets with additional semantic categories by including these types in our dataset as distinct semantic categories. This gave us the following increase in semantic categories: GREC one, EPI 15, ID ten, GENIA nine.

The original GREC corpus contains an exceptionally wide array of semantic categories. While this is desirable for evaluating the performance of our approach under different task settings, the sparsity of the data is a considerable problem; the majority of categories do not permit stable evaluation as they have only a handful of annotations each. To alleviate this problem we used the five ontologies defined in the GREC annotation guidelines[5], collapsing the annotations into five semantic super categories to create a resource we refer to as Super GREC. This preprocessing conforms with how the categories were used when annotating the GREC corpus (Thompson et al., 2009). This resource contains sufficient annotations for each semantic category to enable evaluation on a category-by-category basis. Also, for the purpose of our experiments we removed all "SPAN" type annotations since they themselves carry no semantic information (cf. GREC annotation guidelines).

CALBC CII contains 75,000 documents, which is more than enough for our experiments. In order to maintain balance in size between the resources in our experiments, we sampled a random 5,000 documents and used these as our CALBC CII dataset.

[5] http://www.nactem.ac.uk/download.php?target=GREC/Event_annotation_guidelines.pdf

Name	Abbreviation	Publication
BioNLP/NLPBA 2004 Shared Task Corpus	NLPBA	Kim et al. (2004)
Gene Regulation Event Corpus	GREC	Thompson et al. (2009)
Collaborative Annotation of a Large Biomedical Corpus	CALBC CII	Rebholz-Schuhmann et al. (2010)
Epigenetics and Post-Translational Modifications	EPI	Ohta et al. (2011)
Infectious Diseases Corpus	ID	Pyysalo et al. (2011)
Genia Event Corpus	GENIA	Kim et al. (2011)

Table 4: Corpora used for evaluation

3.3 Corpus Statistics

In this section we present statistics for each of our datasets. For resources with a limited number of semantic categories we use pie charts to illustrate their distribution (Figure 1). For the other corpora we use tables to illustrate this. Tables for the corpora for which pie charts are given has been left out due to space requirements.

The NLPBA corpus (Figure 1a) with 59,601 tokens annotated, covers five semantic categories, with a clear majority of protein annotations. While NLPBA contains several semantic categories, they are closely related, which is expected to pose challenges for disambiguation. This holds in particular for proteins, DNA and RNA, which commonly share names.

Our collapsed version of GREC, Super GREC (see Figure 1b), contains 6,777 annotated tokens and covers a total of six semantic categories: Regulatory Event (GRE), nucleic acids, proteins, processes, living system and experimental. GREC is an interesting resource in that its classes are relatively distinct and four of them are evenly distributed.

CALBC CII is balanced among its annotated categories, as illustrated in Figure 1c. The 6,433 tokens annotated are of the types: proteins and genes (PRGE), species (SPE), disorders (DISO) and chemicals and drugs (CHED). We note that we have introduced lexical resources covering each of these classes (Section 3.1).

For the BioNLP'11 ST resources EPI (Table 5), GENIA (Figure 1d and contains 27,246 annotated tokens) and ID (Table 6), we observe a very skewed distribution due to our decision to include event types as distinct classes; The dominating class for all the datasets are proteins. For several of these categories, learning accurate disambiguation is expected to be very challenging if not impossible due to sparsity: For example, Dehydroxylation in EPI has a single annotation.

Type	Ratio	Annotations
Acetylation	2.3%	294
Catalysis	1.4%	186
DNA demethylation	0.1%	18
DNA methylation	2.3%	301
Deacetylation	0.3%	43
Deglycosylation	0.2%	26
Dehydroxylation	0.0%	1
Demethylation	0.1%	12
Dephosphorylation	0.0%	3
Deubiquitination	0.1%	13
Entity	6.6%	853
Glycosylation	2.3%	295
Hydroxylation	0.9%	116
Methylation	2.5%	319
Phosphorylation	0.9%	112
Protein	77.7%	10,094
Ubiquitination	2.3%	297
Total:		12,983

Table 5: Semantic categories in EPI

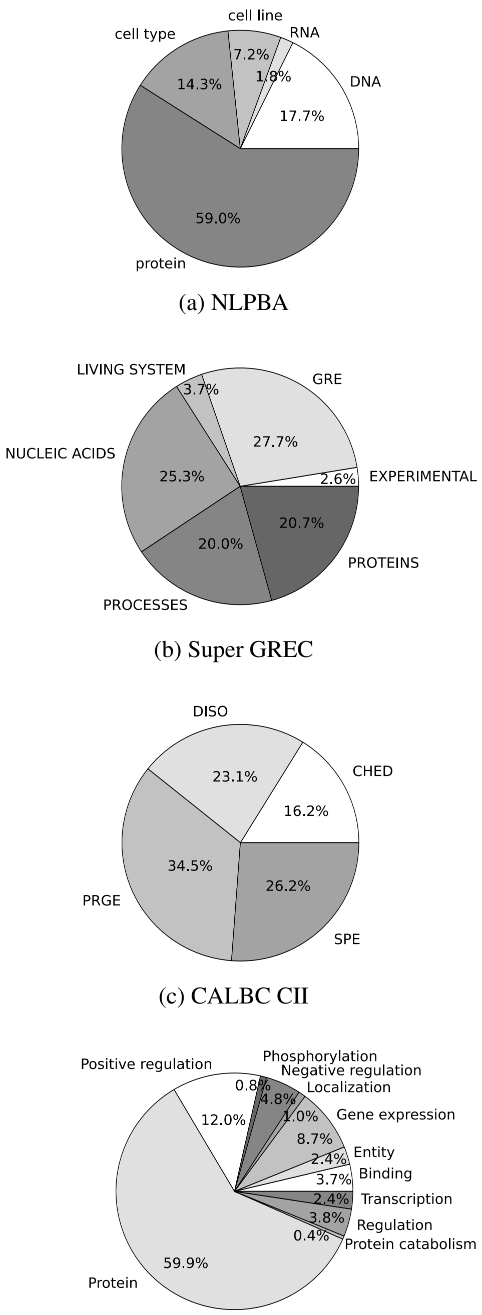
ID is of particular interest since it contains a considerable amount of annotations for more than one physical entity category, including in addition to protein also organism and a minor amount of chemical annotations.

4 Experiments

In this section we introduce our experimental set-up and discuss the outcome of our experiments.

4.1 Experimental Set-up

To ensure that our results are not biased by overfitting on a specific set of data, all data sets were separated into training, development and test sets.



(a) NLPBA

(b) Super GREC

(c) CALBC CII

(d) GENIA

Figure 1: Semantic category distributions

NLPBA defines only a training and test set, GREC and CALBC CII are provided as resources and lack any given division, and for the BioNLP'11 ST data the test sets are not distributed. Thus, we combined all the available data for each dataset and separated the documents into fixed sets with the following ratios: 1/2 training, 1/4 development and 1/4 test.

Type	Ratio	Annotations
Binding	1.0%	102
Chemical	6.8%	725
Entity	0.4%	43
Gene expression	3.3%	347
Localization	0.3%	36
Negative regulation	1.6%	165
Organism	25.5%	2,699
Phosphorylation	0.5%	54
Positive regulation	2.5%	270
Process	8.0%	843
Protein	43.1%	4,567
Protein catabolism	0.0%	5
Regulation	1.8%	188
Regulon-operon	1.1%	121
Transcription	0.4%	47
Two-component-system	3.7%	387
Total:		10,599

Table 6: Semantic categories in ID

We use a total of six classifiers for our experiments. First, a naive baseline (Naive): a majority class voter with a memory based on the exact text of the textual span. The remaining five are machine learning classifiers trained using five different feature sets: gazetteer features constituting strict string matching towards our SimString databases (Gazetteer), SimString features generated from our SimString databases (SimString), the span internal features listed in Table 1 (Internal), the span internal and gazetteer features (Internal-Gazetteer) and the span internal and SimString features (Internal-SimString).

We evaluate performance using simple instance-level accuracy (correct classifications / all classifications). Results are represented as learning curves for each data set.

4.2 Results

From our experiments we find that – not surprisingly – the performance of the Naive, Gazetteer and SimString classifiers alone is comparatively weak. Their performance is illustrated in Figure 2. We can briefly summarize the results for these methods by noting that the SimString classifier outperforms the Gazetteer by a large margin for every dataset.[6] From

[6] Due to space restrictions we do not include further analysis or charts.

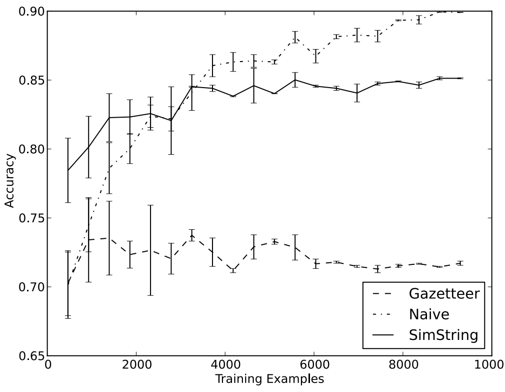


Figure 2: SimString, Gazetteer and Naive for ID

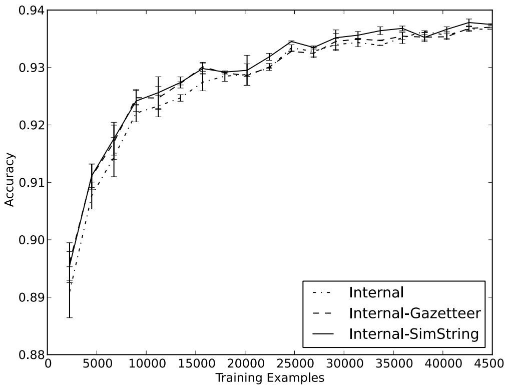


Figure 3: Learning curve for NLPBA

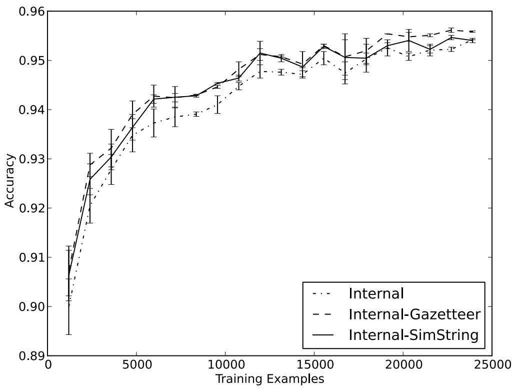


Figure 4: Learning curve for GENIA

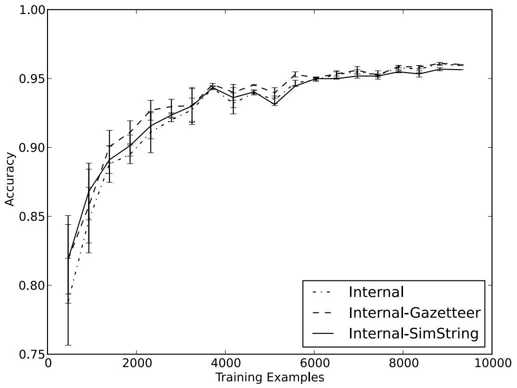


Figure 5: Learning curve for ID

here onwards we focus on the performance of the Internal classifier in combination with Gazetteer and SimString features.

For NLPBA (Figure 3), GENIA (Figure 4) and ID (Figure 5) our experiments show no clear systematic benefit from either SimString or Gazetteer features.

For Super GREC (Figure 6) and EPI (Figure 7) classifiers with Gazetteer and SimString features consistently outperform the Internal classifier, and the SimString classifier further shows some benefit over Gazetteer for EPI.

The only dataset for which we see a clear benefit from SimString features over Gazetteer and Internal is for CALBC CII (Figure 8).

5 Discussion and Conclusions

While we expected to see clear benefits from both using Gazetteers and SimString features, our experiments returned negative results for the majority of the corpora. For NLPBA, GENIA and ID we are aware that most of the instances are either proteins or belong to event trigger classes for which we may not have had adequate lexical resources for disambiguation. By contrast, for Super GREC there are several distinct classes for which we expected lexical resources to have fair coverage for SimString and Gazetteer features. While an advantage over Internal was observed for Super GREC, SimString features showed no benefit over Gazetteer features. The methods exhibited the expected result on only one of the six corpora, CALBC CII, where there is a clear advantage for Gazetteer over Internal and a further clear advantage for SimString over Gazetteer.

Disappointingly, we did not succeed in establishing a clear improvement for more than one of the six corpora. Although we have not been successful in

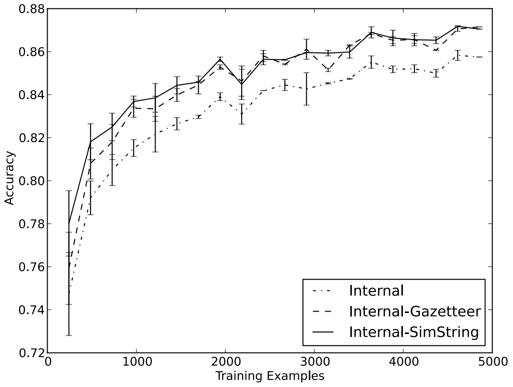


Figure 6: Learning curve for Super GREC

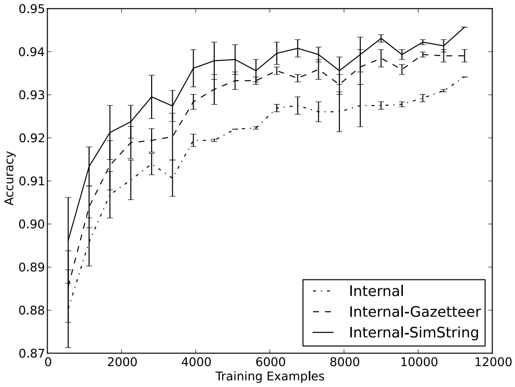


Figure 7: Learning curve for EPI

proving our initial hypothesis we argue that our results calls for further study due to several concerns raised by the results remaining unanswered. It may be that our notion of distance to lexical resource entries is too naive. A possible future direction would be to compare the query string to retrieved results using a method similar to that of Tsuruoka and Tsujii (2003). This would enable us to retain the advantage of fast approximate string matching, thus being able to utilise larger lexical resources than if we were to calculate sophisticated alignments for each lexical entry.

Study of the confusion matrices revealed that some event categories such as negative regulation, positive regulation and regulation for ID are commonly confused by the classifiers. Adding additional resources or contextual features may alleviate these problems.

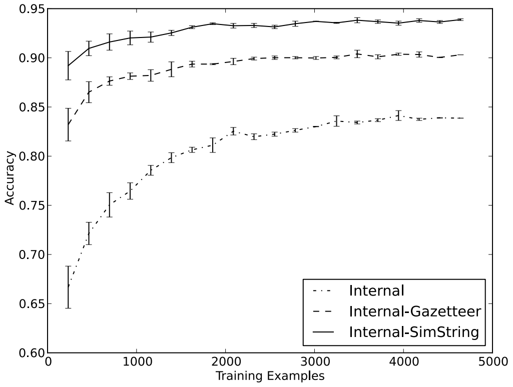


Figure 8: Learning curve for CALBC CII

To conclude, we have found a limited advantage but failed to establish a clear, systematic benefit from approximate string matching for semantic class disambiguation. However, we have demonstrated that approximate string matching can be used to generate novel features for classifiers and allow for the utilisation of large scale lexical resources in new and potentially interesting ways. It is our hope that by making our findings, resources and implementation available we can help the BioNLP community to reach a deeper understanding of how best to incorporate our proposed features for semantic category disambiguation and related tasks.

Our system and collection of resources are freely available for research purposes at http://github.com/ninjin/simsem

Acknowledgements

The authors would like to thank Dietrich Rebholz-Schuhmann and the CALBC organisers for allowing us the use of their data. and Jari Björne for answering questions regarding the Turku Event Corpus. We would also like to thank the anonymous reviewers and Luke McCrohon for their insightful and extensive feedback, which has considerably helped us to improve this work. Lastly the first author would like to thank Makoto Miwa and Jun Hatori for their timely and helpful advice on machine learning methods.

This work was supported by the Swedish Royal Academy of Sciences and by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25.

E. Beisswanger, M. Poprat, and U. Hahn. 2008. Lexical Properties of OBO Ontology Class Names and Synonyms. In *3rd International Symposium on Semantic Mining in Biomedicine*.

J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 28–36. Association for Computational Linguistics.

O Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

M.F.M. Chowdhury and A. Lavelli. 2010. Disease Mention Recognition with Specific Features. *ACL 2010*, page 83.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

M. Gerner, G. Nenadic, and C.M. Bergman. 2010. LINNAEUS: A species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85.

K.M. Hettne, R.H. Stierum, M.J. Schuemie, P.J.M. Hendriksen, B.J.A. Schijvenaars, E.M. Mulligen, J. Kleinjans, and J.A. Kors. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22):2983.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 70–75.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Jin-Dong Kim, Yue Wang, Toshihasi Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

R. Leaman and G. Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663. Citeseer.

D. Maglott, J. Ostell, K.D. Pruitt, and T. Tatusova. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(suppl 1):D54.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China, August.

M.F. Porter. 1993. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

D. Rebholz-Schuhmann, A.J.J. Yepes, E.M. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn. 2010. CALBC silver standard corpus. *Journal of bioinformatics and computational biology*, 8(1):163–179.

Y. Sasaki, Y. Tsuruoka, J. McNaught, and S. Ananiadou. 2008. How to make the most of NE dictionaries in statistical NER. *BMC bioinformatics*, 9(Suppl 11):S5.

L. Shi and F. Campagne. 2005. Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC bioinformatics*, 6(1):88.

L. Tanabe, N. Xie, L. Thom, W. Matten, and W.J. Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(Suppl 1):S3.

P. Thompson, S.A. Iqbal, J. McNaught, and S. Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10(1):349.

M. Torii, Z. Hu, C.H. Wu, and H. Liu. 2009. BioTagger-GM: a gene/protein name recognition system. *Journal of the American Medical Informatics Association*, 16(2):247.

Y. Tsuruoka and J. Tsujii. 2003. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 41–48. Association for Computational Linguistics.

Yue Wang, Jin-Dong Kim, Rune Saetre, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC Bioinformatics*, 10(1):403.

C.H. Wu, L.S.L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R.S. Ledley, B.E. Suzek, et al. 2003. The protein information resource. *Nucleic Acids Research*, 31(1):345.