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

Pontus Stenetorp[*] **Sampo Pyysalo**[†‡] **Sophia Ananiadou**[†‡] **and Jun'ichi Tsujii**[$]
[*]Aizawa Laboratory, Department of Computer Science, The University of Tokyo, Tokyo, Japan
[†]School of Computer Science, University of Manchester, Manchester, UK
[‡]National Centre for Text Mining, University of Manchester, Manchester, UK
[$]Microsoft Research Asia, Beijing, People's Republic of China
{pontus,smp}@is.s.u-tokyo.ac.jp
sophia.ananiadou@manchester.ac.uk
jtsujii@microsoft.com

Abstract

In this study we investigate how semantic category disambiguation can be used to support other Natural Language Processing tasks and annotation efforts. While previous research has mostly cast semantic category disambiguation purely as a classification task, we propose a task setting analogous to dynamic width beam search that allows for multiple semantic categories to be suggested while aiming to minimise the number of suggestions and maintain high recall. We base our approach on a recent machine learning-based system and evaluate it on six recently introduced corpora, one incorporating as many as 17 semantic categories, our system performs in the recall range of 98.6% to 99.5% while keeping the average number of semantic categories proposed in the range of 1.3 to 2.0. The level of performance suggests that the system is adequate to meet the human requirements of human annotators and could successfully be used for annotation support. The introduced system and all related resources are freely available for research purposes at: https://github.com/ninjin/simsem

1 Introduction

Semantic category disambiguation is a key sub-task of several core problems in Natural Language Processing (NLP) such as co-reference and coordination resolution. It is of particular importance for Named Entity Recognition (NER) which is concerned with the detection and demarcation of semantic category mentions in text. Conceptually, NER involves two sub-tasks that must be solved: detecting entity mentions and determining to which semantic category a given mention belongs.

An example of a practical application which requires semantic category disambiguation is the annotation of textual spans or documents. For example when assigning labels such as those of ICD-10 (Resnik et al., 2006) or producing annotations to train information extraction systems (Verspoor et al., 2009). For any assignment task there are cognitive limitations on the number of distinct categories a human annotator can process before falling victim to degrading performance. Thus an automated system could assist annotators by limiting the number of categories presented to the user, excluding those which are clearly irrelevant.

However, such a system would be subject to much scrutiny and must therefore have a very high degree of recall not to cause frustration over the system itself while at the same time limit the number of categories presented to the highest degree possible, even when the amount of training data is sparse.

2 Previous Work

Although semantic category disambiguation is central to NER, there have been relatively few in-domain studies investigating semantic category disambiguation as a stand-alone task. However, recently a few publications has investigated this task in isolation.

Cohen et al. (2011) presented a fast and reliable approach to associating a given textual span to one or several ontologies. The method was based on a set of manually crafted rules and achieved a macro-level accuracy ranging from 77.1% to 95.5% when

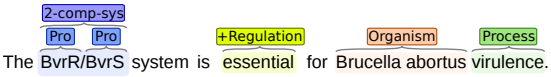


Figure 1: Example of typed text-bound annotations from Pyysalo et al. (2011)

determining from which ontology a given annotation in a corpus was derived from.

In recent work (Stenetorp et al., 2011b) we introduced a machine learning-based method that employed approximate string matching of textual spans to several large-scale lexical resources for the purpose of semantic category disambiguation. Employing lexical resources such as dictionaries covering specific semantic categories is commonplace for state-of-the-art NER systems (Torii et al., 2009; Ratinov and Roth, 2009), but approximate string matching was a novel aspect. We evaluated the method on several datasets and achieved results ranging from 85.9% to 95.3% in macro-level accuracy. However, we failed to establish a clear systematic benefit of approximate, as opposed to strict, string matching for all datasets.

Since our aim is to evaluate the performance of semantic category disambiguation for assisting other tasks such as annotation, the approach of Cohen et al. (2011) has two limitations. It assumes that the semantic categories are defined by ontology resources and therefore it is unable to adapt to an annotation target that doesn't match the ontological resources available. Secondly, unlike Stenetorp et al. (2011b) their approach does not provide ranking or classification confidence. Since this makes it less suitable in a setting where the number of suggested categories are to be kept to a minimum, as is the case for annotation support, we choose to extend our previous system and evaluate its applicability to support other NLP tasks.

3 Methods

3.1 Task Setting

Given a text and a continuous textual span, we classify and assign one category out of several semantic categories from a fixed set. Figure 1 illustrates the style of text-bound annotations[1] and the possibility of overlapping spans with different semantic categories.

While in Stenetorp et al. (2011b) we cast semantic category disambiguation purely as a classification task and Cohen et al. (2011) as a classification task with multiple possible correct labels, we suggest to take a different perspective. We propose a task setting where we allow the method to return multiple suggestions for a given annotation. This setting is analogous to beam search with a dynamic width beam (Ney et al., 1992) in that it must maintain high recall while keeping the number of suggestions at a minimum and is also conceptually similar to scoring methods such as Mean Reciprocal Rank from the field of Information Retrieval. This fulfils our goal of capturing the cognitive burden on a human annotator having to determine the correct answer among multiple suggestions and also captures how well a method can estimate it's own performance when passing on suggestions to another system.

4 Experimental Set-up

4.1 Metrics

We train our model and produce learning curves with data points using: $[5\%, 10\%, \ldots, 100\%]$ of the training data respectively. For each data point we measure the ambiguity which is the average number of suggested categories and recall by the number of correct categories left out by the system. At each data point we take several random samples of the current data size and use the mean of the performance over the samples. Results for each metric are provided as the mean of the data points of the learning curve (analogous to the Area Under the Curve).

4.2 Models

In parallel with our experiments we constructed and evaluated several models in addition to those previously published for our system. For details regarding these models and their performance we refer the reader to Stenetorp et al. (2011a).

For the task setting we introduced in Section 3.1 we need to adapt our system so that it is capable of determining how many categories to propose for a given annotation. Our machine learning method provides probabilistic output which can be used to de-

[1] Visualised using: https://github.com/TsujiiLaboratory/stav

Data set	μAmb.	Amb.	μRecall	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 1: Performance by ambiguity level/reduction (Amb.) and recall for mean (μ) over the learning curve and when all training and development data was used as training data

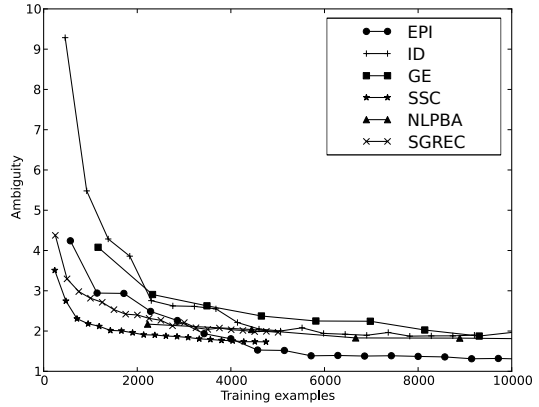
termine the confidence the system has for a given annotation and we can then select a confidence threshold limiting the number of suggestions. The system is set to return the smallest collection of suggestions so that the sum of the confidence for all the suggestions returned reach the given threshold, this will return from 1 to the total number of categories in the dataset suggestions. For example: For a textual span we have the categories and confidences [PROTEIN 90%, ORGANISM 6%, CHEMICAL 4%] and a confidence threshold of 95%. The system would present PROTEIN and ORGANISM since given the confidence threshold we can discard CHEMICAL at the risk of dropping recall but with the benefit of reducing the ambiguity of the output.

4.3 Corpora

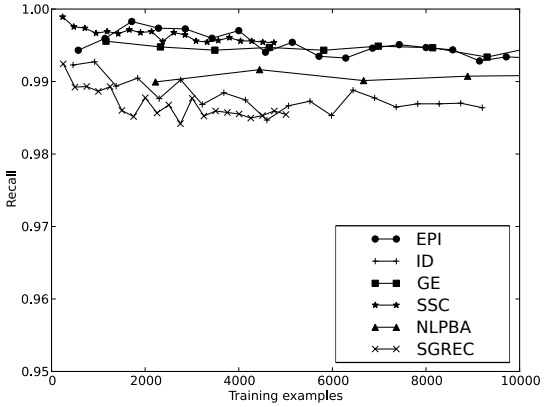
We evaluate our proposed model on the six datasets used in Stenetorp et al. (2011a), due to space requirements we refer the reader to the previous publication for additional information on the datasets used. All the datasets were randomly separated into training, development and test sets consisting of 1/2, 1/4 and 1/4 of the annotations respectively. The test set was kept hidden during development and was only used to generate the final results prior to submitting the publication.

5 Results and Discussion

We performed a small set of experiments on the development set and found that a confidence threshold of 99.5% and INTERNAL-SIMSTRING with a cosine threshold of 0.4 (Stenetorp et al., 2011a) to be the best model and decided to use it for our final experiments.



(a) Ambiguity



(b) Recall

Figure 2: Learning curves for ambiguity and recall

In Figure 2a and Figure 2b we can see the lower end of the learning curves for ambiguity and recall for all datasets. When it comes to the level of ambiguity and recall our results look very promising. While at first this may be surprising it is not unintuitive, the system has simply prioritised recall over ambiguity since the optimisation target for our model is accuracy.

For Figure 2a we see that the ambiguity quickly drops to a manageable level, as per Table 1 the reduction in the number of semantic categories is on average at least 50% for each dataset. Our most impressing results are for EPI where for even the smallest training size data point the number of categories are reduced from 17 to $\sim$4.5, using all the training

data the output only exposes 10.6% of the 17 categories which is a considerable reduction. For all datasets the results are achieved while recall stays consistently at ∼99% which is a tolerable level.

It should also be noted that although our results are promising, whether or not they are sufficient for a human annotator or to support another NLP system is a different matter. In particular for human annotators it may be necessary to weight the recall versus ambiguity very carefully, perhaps even to suit each annotator's needs, which can easily by controlled for our proposed method by adjusting the confidence threshold. A human annotator may tolerate a lower level of recall in favour of lower ambiguity, a coreference resolution system would be more likely to prefer close to perfect recall at the cost of higher ambiguity.

6 Conclusions and Future Research

In this paper we have investigated ways in which semantic category disambiguation can be used to support other NLP tasks and act to support a human annotator. Having introduced a task suitable for measuring a systems performance for such a setting and adapted an existing method for it we find that our recall is ∼99% for all datasets while we are capable of reducing the average number of suggested categories to at least 50% of the original number of categories. A level which is easily manageable by human annotators.

For future research we intend to incorporate our system into an annotation tool and measure the impact on annotator performance if we reduce the number of categories presented to the annotator, possibly even requiring no annotator feedback if only one category is enough to meet the confidence threshold. We will also seek to investigate if the system can be used as a safe-guard to prevent annotation mistakes which are simple slips by the annotator and is a common source of errors for manually annotated corpora. This can be done by raising a warning if the annotators judgement largely disagrees with that of the systems, either interactively or as a post-processing step to verify annotation quality.

Our system, additional results and related resources are freely available for research purposes at: https://github.com/ninjin/simsem

Acknowledgements

We would like to thank the anonymous reviewers for improving the paper by providing several helpful comments and references.

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).

References

K.B. Cohen, T. Christiansen, W. Baumgartner Jr., K. Verspoor, and L. Hunter. 2011. Fast and simple semantic class assignment for biomedical text. In *Proceedings of BioNLP 2011 Workshop*, pages 38–45.

H. Ney, D. Mergel, A. Noll, and A. Paeseler. 1992. Data driven search organization for continuous speech recognition. *IEEE Transactions on Signal Processing*, 40(2):272–281.

S. Pyysalo, T. Ohta, R. Rak, D. Sullivan, C. Mao, C. Wang, B. Sobral, J. Tsujii, and S. Ananiadou. 2011. Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL 2009*, pages 147–155.

P. Resnik, M. Niv, M. Nossal, G. Schnitzer, J. Stoner, A. Kapit, and R. Toren. 2006. Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding. In *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings*.

P. Stenetorp, S. Pyysalo, S. Ananiadou, and J. Tsujii. 2011a. Investigating approaches to semantic category disambiguation using large lexical resources and approximate string matching. In *IPSJ SIG Notes (to appear)*. Information Processing Society of Japan (IPSJ).

P. Stenetorp, S. Pyysalo, and J. Tsujii. 2011b. Simsem: Fast approximate string matching in relation to semantic category disambiguation. In *Proceedings of BioNLP 2011 Workshop*, pages 136–145.

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.

K. Verspoor, K.B. Cohen, and L. Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, 10(1):183.