

# Predicting Domain Specific Entities with Limited Background Knowledge

Christopher Thomas and Amit P. Sheth

LSDIS lab, Computer Science Department, University of Georgia, Athens

[chaos@uga.edu](mailto:chaos@uga.edu), [amit@cs.uga.edu](mailto:amit@cs.uga.edu)

## Abstract.

This paper proposes a framework for automatic recognition of domain-specific entities from text, given limited background knowledge, e.g. in form of an ontology. The algorithm exploits several lightweight natural language processing techniques, such as tokenization and stemming, as well as statistical techniques, such as singular value decomposition (SVD) to suggest domain relatedness of unknown entities.

## 1 Introduction

Great progress has been made in the classification of documents in the biomedical field. PubMed[7] categorizes publications according to concepts in the MeSH[8] hierarchy. Immense human efforts went into these endeavors. MeSH was created manually by experts and the documents are manually classified according to major topics. Assuming, however, that the knowledge contained in these publications is to be used to help validating hypotheses using computational methods, not only single document need to be categorized and annotated, but also entities and relationships used and described in the publications. In order to annotate documents extensively, terms and phrases have to be identified and disambiguated to find the correct annotation. A first step in this direction is the identification of biomedical terms or entities.

Different techniques have been proposed for entity recognition from text. Many assume extensive background knowledge[2], others use machine learning techniques, such as Hidden Markov Models[12] or SVMs[9]. See[10] for an extensive review of the related work.

This paper proposes a framework for entity recognition based on several lightweight NLP techniques and unsupervised statistical processing with limited background knowledge. The NLP steps include domain-specific tokenization, POS tagging and chunk parsing. The statistical processing is done by computing the pair wise distances between vector space representations of terms in a TFIDF term-document matrix, optionally after performing Singular Value Decomposition (SVD). An advantage of this largely unsupervised technique over traditional Machine

Learning techniques is that it can be applied to various domains simply by changing the source of the background knowledge, i.e. the ontology or ontologies fed into the system. However, contrary to many ontology based Entity Recognition systems, the ontology serves not as the dictionary, but as a means to identify terms that are not in the ontology itself, but belong to the same domain.

The remainder of the paper is structured as follows. Section 2 describes the architecture of the framework. Section 3 gives an extensive performance evaluation of various possible configurations. Section 4 concludes the paper and gives an outlook to future work.

## 2 System Description

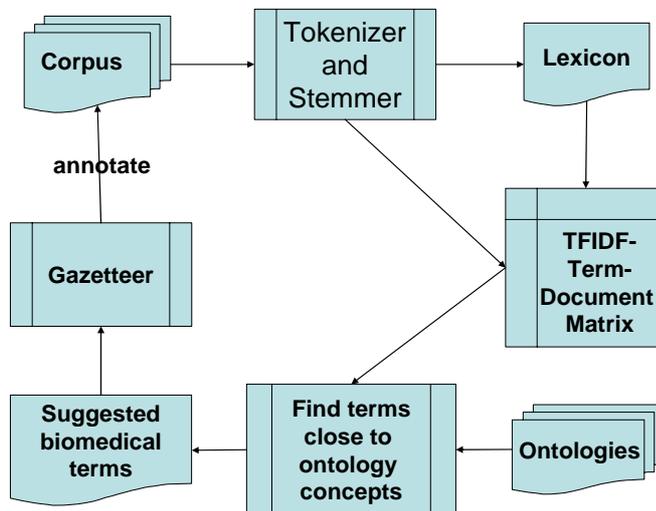


Figure 1: The basic system architecture

### 2.1 Term Extraction, Lexicon and Matrix Building

In the following, the base algorithm is described, as depicted in Figure 1, which is later modified to allow for different evaluation scenarios. In a multi-step process, all terms in the document corpus are tokenized, stemmed and extracted into a lexicon, which defines the columns of a document-term matrix. The raw counts of the words that appear in each document are then filled into the rows of the matrix. After this step, the matrix is recomputed according to the TF-IDF formula. This assures that the importance of the term for the document is taken into account, not its pure frequency. The columns represent word stems that can possibly map to several terms in the

document corpus. The lexicon contains a mapping from word stems to original appearances, which are then used for evaluation.

For the non-SVD evaluation, the resulting matrix is transposed into a term-document matrix, in which a single row represents the occurrence of one term in all documents.

When SVD is used, the document-term matrix is transformed using Singular Value Decomposition, resulting in three matrixes U, S and V. The product of S and V is a representation of the terms in the transformed vector space.

The corpus lexicon itself can be post-processed in different ways to allow for a bias towards a given domain. For example, all terms that are known not to be domain related can be removed for faster processing, since the matrix size will be heavily reduced. This might have negative impact on the statistical analysis, though, because domain terms tend to occur in conjunction with non-domain terms. For the evaluation section, this has not been applied, because the outcome of suggested domain terms is evaluated with respect to a domain dictionary.

## 2.2 Ontology Comparison

The next step is the computation of terms in the lexicon that are statistically close to terms used in the ontologies of choice. Those vectors in the matrix representing terms that are both in the dictionary and in the ontology are chosen for comparison. Then, the pair-wise cosine-distance between each other term and the ontology terms is computed and the min/average distance is kept for each of these terms. Finally, the terms are sorted according to their distance from the ontology terms. Presumably, terms with less distance to the ontology terms are more likely to belong to the domain of the ontology than terms that are more distant. Intuitively, choosing a cutoff point N that considers fewer terms as being domain-related gives better precision, but less recall than one that considers more terms. An ideal point N will need to be determined empirically, but the results so far suggest that 10% of the lexicon size is a good measure. In the evaluation, charts are drawn that show the precision for all possible cutoff points. It is important to note that in all test runs, the terms in the ontology constituted approximately 1% of all corpus terms.

## 2.2 Corpus annotation

Using the terms in the comparison ontologies and the suggested domain terms as dictionaries, the GATE[3] gazetteer is run to annotate the documents in the corpus. For an XML output, tags are placed around each identified concept. The gazetteer distinguishes between concepts that are known with certainty, because they are found in the ontologies and the suggested biomedical entities, which are only assumed to be domain entities. Known entities are annotated with their ontology concept name, while the suggested entities are annotated with the tag *suggestedBioEntity*. These terms can be subject to further verification.

### 3 Evaluation

#### 3.1 Evaluation with respect to the SPECIALIST lexicon

The Entity Recognition algorithm was tested on several corpora of publication abstracts taken from PubMed under the heading *Glycosylation* in order to have a corpus that is close to Glycomics, the domain of the GlycO[11] ontology. GlycO has been developed as a focused domain ontology for the glycobiology field. The domain that it describes is comparably narrow and the description is very accurate. To verify the claim that the algorithm finds terms of the same domain with high accuracy, the upper level ontology SUMO[4] is used as a comparison. Precision measures are reached by comparing the lists of suggested biomedical entities to terms in the SPECIALIST lexicon[5]. One problem with this lexicon is, that it does not only contain terms that solely exist in the biomedical domain, so a comparison ontology from a broader domain can also yield good results, even though the matches actually occur in the wrong domain. One evaluation strategy that tries to avoid this problem is to remove all the terms from the SPECIALIST lexicon, that are also in WordNet[6]. However, WordNet also contains some biomedical terms and hence the precision will be lower for all comparison ontologies. Thus, these charts need to be seen only as a relative comparison for the precision using different comparison ontologies, not as absolute values. Another drawback of the automatic evaluation is that many of the terms that are not found are actually domain terms, e.g. gene names or compound terms that do not appear in the SPECIALIST lexicon. The following table contains a list of the terms that have not been found in SPECIALIST, in the order of their likelihood of being biomedical domain terms.

manno, deoxy, phthalimido, mannopyranosyl, gave, thio, u266, volcanii, nalm, fgfr1, deallylation, desilylation, spectra, jok, cd44s, mannospyranoside, uea, alpha2, pgii, vaa, dlif, tetraisopropyl, disiloxane, eb6, cd44v, bl6, ssbetagly, tf1, cld, mice, rldti, oxpc, ova, celb, cruzi, beta1, igan, hydrolysing, epcr, hla, xg9, phosphopolyisoprenols, pp55, decanoylamino, pp36, 3gt, desorption, 3gal, brs, brucei, diyl, htg, cd52, methylamino
--

**Table 1:** Entities not recognized by SPECIALIST

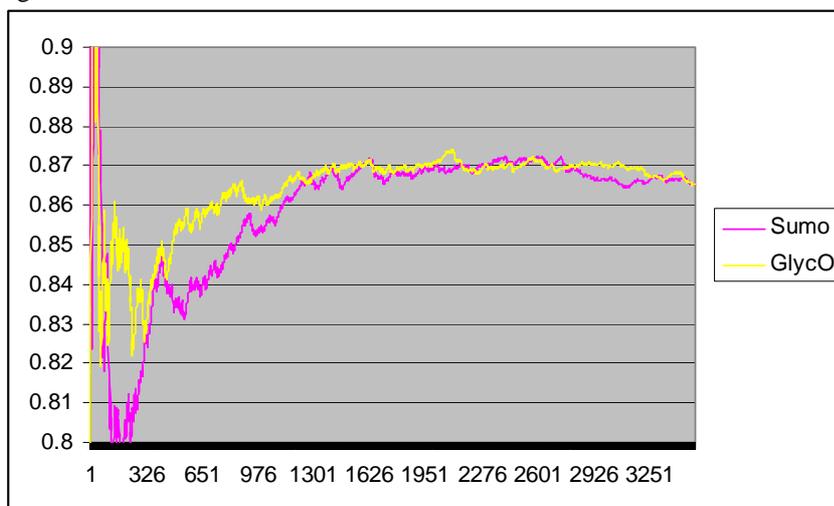
This list contains significantly more biomedical terms than non-biomedical and needs to be evaluated with respect to other biomedical dictionaries and databases. This will be addressed in future work. For this paper, it is important to see the results in light of the fallibility of the SPECIALIST lexicon. It still serves well to distinguish the performance of the algorithm using domain-specific vs. non-domain-specific comparison ontologies.

The following charts show the precision distribution, when using both the SUMO as well as the GlycO ontology for comparison. In the case of **Figure 2**, the lexicon

has not been processed using any background knowledge after tokenization and stemming. Intuitively, since the full list of suggested entities is taken into account for evaluation, the precision of both cases is converging. The first values are naturally jumpy, because the denominator in the precision formula has more impact, but it is clearly visible that for the first 300 terms identified the GlycO comparison gives a higher precision than the SUMO comparison. Precision is computed according to the following formula:

$$\text{Precision}(\text{position } N) = \frac{\# \text{terms identified in SPECIALIST until position } N}{N} \quad (1)$$

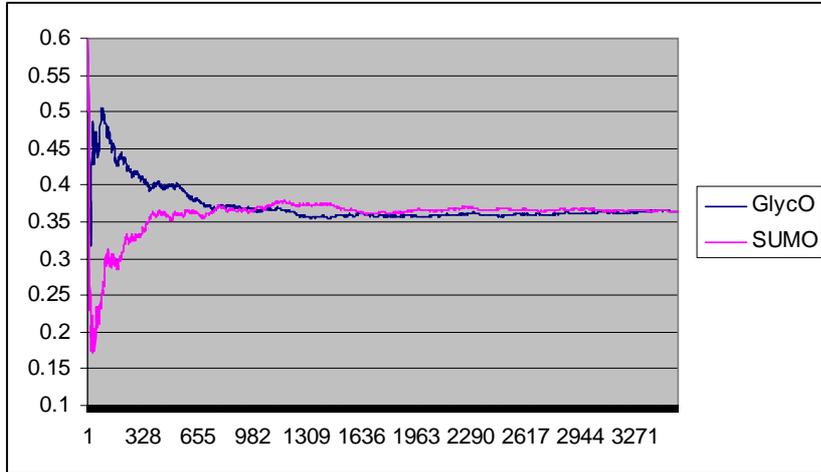
Depending on the number of distinct terms in the corpus, the number of terms in the dictionary varies. The following 2 charts represent evaluations of a run with 200 document abstracts, resulting in more than 3800 distinct stemmed term. **Figure 2** suggests that the average percentage of biomedical entities in the corpus, measured by the terms in SPECIALIST, is more than 86%. The precision up to roughly the 1200<sup>th</sup> identified term actually suggests that the algorithm performs under the baseline. Looking at the terms in **Table 1** and the evaluation in **Figure 2** indicates, however, that the first suggested domain entities often constitute actual domain entities that are not part of the SPECIALIST lexicon. The difference between using the domain ontology GlycO and the upper ontology SUMO is visible, even though not significant.



**Figure 2:** Precision distribution when SVD has not been applied and the full SPECIALIST lexicon has been used for evaluation

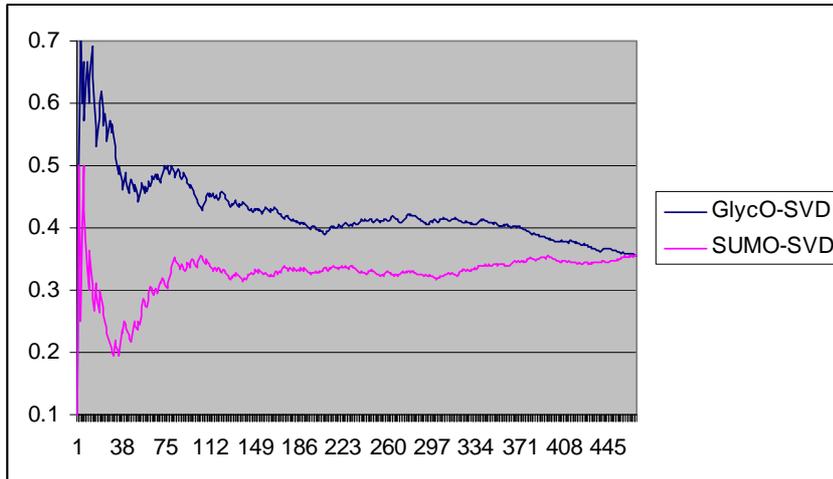
The difference becomes more apparent, when the 2 resulting sets of domain entity suggestions are evaluated with respect to the reduced SPECIALIST lexicon, i.e. terms that are only in SPECIALIST, but not in WordNet. As expected, the precision is overall less, because the identified terms that are in WordNet are not counted as hits. In **Figure 3** it becomes apparent, though, that using GlycO as a comparison yields the

predicted results, with accurately predicted domain terms among the first N suggested domain entities.



**Figure 3:** precision using the modified SPECIALIST lexicon as reference

. Applying singular value decomposition to the term-document matrix yields the best results. **Figure 4** shows the precision of the first 500 suggested domain terms, computed from an SVD-transformed matrix. The algorithm transforms slightly better than in the non-transformed case. Especially the first 40 matches are very accurate and clearly distinguished from the matches that are achieved using SUMO as a comparison ontology..



**Figure 4:** precision of the SVD-based comparison using the modified SPECIALIST as reference

### 3.2 Evaluation of the annotation

The following shows an excerpt of a PubMed abstract annotated with suggested biomedical entities. The tags for *alpha*, *beta* and *deoxy* represent concepts that appear in the GlycO ontology, the *suggestedBioEntity* tags represent terms identified by the entity recognition algorithm.

```
[...]non- <suggestedBioEntity>fucosylated</suggestedBioEntity> core structure of
<suggestedBioEntity>xylose</suggestedBioEntity> -containing
<suggestedBioEntity>carbohydrate</suggestedBioEntity> chains from N-
<suggestedBioEntity>glycoproteins</suggestedBioEntity> . The synthesis is reported of
<suggestedBioEntity>methyl</suggestedBioEntity> 2-
<suggestedBioEntity>acetamido</suggestedBioEntity> -4-O-[2-
<suggestedBioEntity>acetamido</suggestedBioEntity> -2- <deoxy>deoxy</deoxy> -O-(3,6-
di -O- <alpha>alpha</alpha> -D-
<suggestedBioEntity>mannopyranosyl</suggestedBioEntity> -2-O- <beta>beta</beta> -D-
<suggestedBioEntity>xylopyranosyl</suggestedBioEntity> - <beta>beta</beta> -D-
<suggestedBioEntity>mannopyranosyl</suggestedBioEntity> )- <beta>beta</beta> -D-
<suggestedBioEntity>glucopyranosyl</suggestedBioEntity> ]-2- <deoxy>deoxy</deoxy> -
<beta>beta</beta> -D- <suggestedBioEntity>glucopyranoside</suggestedBioEntity> (4)
and <suggestedBioEntity>methyl</suggestedBioEntity> 2-
<suggestedBioEntity>acetamido</suggestedBioEntity> -4-O-[2-
<suggestedBioEntity>acetamido</suggestedBioEntity> -2- <deoxy>deoxy</deoxy> -4-O-
(3,6- di -O- <alpha>alpha</alpha> -D-
<suggestedBioEntity>mannopyranosyl</suggestedBioEntity> -2-O- <beta>beta</beta> -D-
<suggestedBioEntity>xylopyranosyl</suggestedBioEntity> - <beta>beta</beta> -D-
<suggestedBioEntity>mannopyranosyl</suggestedBioEntity> )- <beta>beta</beta> -D-
<suggestedBioEntity>glucopyranosyl</suggestedBioEntity> ]-2- <deoxy>deoxy</deoxy> -
6- O- <alpha>alpha</alpha> -L-
<suggestedBioEntity>fucopyranosyl</suggestedBioEntity> - <beta>beta</beta> -D-
<suggestedBioEntity>glucopyranoside</suggestedBioEntity> (5), which represent the
<suggestedBioEntity>invariant</suggestedBioEntity> [...]
```

Most biomedical entities are correctly identified. The gazetteer was given the 10% of the corpus terms that were closest to the terms in the GlycO ontology for annotation.

## 4 Conclusion and Future Work

The goal of this work was to develop a domain independent Entity Recognition algorithm using lightweight NLP techniques and limited background knowledge in form of domain ontologies. We have presented a framework for identification of domain specific entities, with special consideration of the biomedical domain. The research has shown that it is possible to predict domain specific entities with high likelihood, given limited background knowledge. While this is good enough for information retrieval in a given domain, annotation of entities for the semantic web needs certainty. For this case, the framework can be used as a preprocessing step to place the identified entities in the domain under consideration, for further evaluation

by domain experts or using domain specific lexica. The advantage is still, that fewer terms need to be considered than in case of not pre-processing the raw text.

The future work will include using available agents for Entity Recognition. These agents are usually domain dependent. For the focus domain, agents for the recognition of gene names and other biomedical entities will be deployed. Especially in the complex carbohydrate domain, many terms are compounds that describe chemical structures. Special tokenizers and parsers are developed to split those compound terms into sets of atomic entities. The distance between the terms in the comparison ontology and the corpus terms provide only a relative likelihood of the term belonging to the domain. Further analysis needs to show if this can be translated into an absolute probability.

## 5 References

- [1] Cunningham H., Maynard D., Bontcheva K. and Tablan V., *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. In proc. of the 40<sup>th</sup> Anniversary Meeting of the Association for Computational Linguistics, 2002
- [2] Hanisch, D., Fundel, K., Mevissen, H.-T. et al. *ProMiner: Organism-specific protein name detection using approximate string matching*, in Proceedings of the EMBO workshop BioCreative: Critical Assessment for Information Extraction in Biology, 2004.
- [3] <http://gate.ac.uk/>
- [4] <http://ontology.teknowledge.com/>
- [5] <http://specialist.nlm.nih.gov/>
- [6] <http://wordnet.princeton.edu/>
- [7] <http://www.ncbi.nlm.nih.gov/entrez/>
- [8] <http://www.nlm.nih.gov/mesh/>
- [9] Kazama, J., Makino, T., Ohta, Y, Tsujii, J. *Tuning support vector machines for biomedical named entity recognition*, Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3, July 2002
- [10] Leser, U and Hakenberg, J. *What makes a gene name? Named entity recognition in the biomedical literature*, in Briefings in Bioinformatics, 6(4), 357–369. December 2005
- [11] Thomas, C, Sheth, A, and York, W. *Modular Ontology Design Using Canonical Building Blocks in the Biochemistry Domain*. To appear in the proceedings of the FOIS conference 2006
- [12] Zhao, S.J. *Name Entity Recognition in Biomedical Text using a HMM model*, In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), Geneva, Switzerland..