Lexical Properties of OBO Ontology Class Names and Synonyms

Elena Beisswanger

Michael Poprat

Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab Friedrich-Schiller-Universität Jena D-07743 Jena, Germany

{elena.beisswanger|michael.poprat|udo.hahn}@uni-jena.de

Abstract

While the Open Biomedical Ontologies (OBO) are successfully used for manual database annotation purposes, their usefulness for automatic text mining remains to be shown. Crucial for OBO's suitability for natural language processing applications is the nature of the class names and synonyms provided and the way they are referred to or literally appear in the literature. Accordingly, our study investigates the lexical properties and the semantic ambiguity of these terms. In particular, we identify the number of OBO classes that can be recognized in two corpora by means of these terms, considering one corpus of full texts documents taken from PUBMED Central and one of MEDLINE abstracts. We found that 15% of all OBO classes could be identified in the MEDLINE corpus and 9% in the PUBMED Central corpus by caseinsensitive string matching, including term variants and stemmed forms of terms. Interestingly enough, only nine out of 80 OBO ontologies account for 80% of the OBO classes that we were able to find.

1 Introduction

The Open Biomedical Ontologies (OBO) library¹ is a collection of publicly available biomedical ontologies hosted by the U.S. National Center for Biomedical Ontologies (NCBO). OBO ontologies cover different subdomains of biology and biomedicine, amongst others the anatomy of different model organisms, biomedical processes, molecular functions of gene products, sequence features, chemicals and experimental methods. The ontologies have been developed as controlled

vocabularies for various data management tasks. For example, the Gene Ontology (GO),² the most prominent one of the OBO ontologies, was created for the functional annotation of genes and gene products. The aim of using shared controlled vocabularies such as the GO is to facilitate the interoperability of different but related biomedical databases across species.

OBO ontologies hold domain-specific knowledge in a structured way by using hierarchically organized classes and additional semantic relations. Classes come with a class name and are often supplemented with synonyms and textual definitions. While typically the class name is unambiguous and self-explaining, the synonyms are supplied to reflect the natural language use in documents and thus tend to be ambiguous. Besides the hierarchy defining *is-a* relation, many OBO ontologies provide complementary semantic relations (such as *part-of* and *develops-from*) to express complex domain-specific interdependencies (e.g., "cellular membrane" *part-of* "cell", "mature T cell").

The OBO ontologies provide a huge amount of domain-specific vocabulary in terms of class names and their synonyms, which in their entirety we refer to as OBO *terms*. Although it is well known that natural language processing (NLP) may heavily benefit from access to biomedical domain knowledge (Spasic et al., 2005), re-use of OBO for NLP is rare. This might be due to the fact that the OBO terms are suspected to be rather artificial, utterly long and complex – taking the lexical properties of GO class names (McCray et al., 2002; Ogren et al., 2004) *pars pro toto* for OBO terms.

http://obofoundry.org/

²http://www.geneontology.org/

In order to find out whether this assumption is justified or not we here scrutinize on natural language properties of OBO terms. We do so by looking for matches of OBO terms in two natural language corpora, subsets of MEDLINE and PUBMED Central. We also incorporate another terminological umbrella system, the Unified Medical Language System (UMLS), that enjoys much greater acceptance in the biomedical NLP community than the OBO ontologies.

2 Related Work

Several studies focusing on the lexical nature of terms in domain-specific terminologies and ontologies have already been carried out. One of their main intentions is to find ways how these terms can be arranged in domain-specific lexicons that are easy to use by NLP engines.

Verspoor (2005) constructed a semantic lexicon based on terms that occur in both, the UMLS Metathesaurus³ and the UMLS Specialist Lexicon.⁴ Several matching strategies were applied to detect the UMLS terms in a domain-specific text corpus. Verspoor found a lexical overlap for 77% of the tokens in the corpus, though regarding the set of all different tokens (i.e., types) in the corpus only 3% were matched. The study did not explore whether the meaning of the terms found in a text really corresponded to the meaning of the targeted terms in the merged lexicon, i.e., the resolution of semantic ambiguity is left as an open issue.

McCray et al. (2001) also evaluated UMLS Metathesaurus terms regarding their usefulness for NLP. They identified different string properties that allow to predict how likely it is that a given term can be found in a domain-specific text corpus. In a follow-up study they explored the lexical properties of Go terms. In particular, they determined the number of Go terms that appear in the UMLS Metathesaurus as well, checked whether the terms fulfilled certain lexical properties that indicate their 'wellformedness' for NLP and looked for Go terms in a domain-specific text corpus (McCray et al., 2002). Another stream of work on Go terms investigates their compositionality (see, e.g., Ogren et al. (2004)) rather than their lexical features in relation to NLP tasks.

Our study resembles the one by McCray et al. (2002) but extends its scope of analysis from the GO to the whole of OBO. In addition, we go beyond the work of McCray et al. and Verspoor by investigating the semantic ambiguity of OBO ontology terms and collecting preliminary evidence for the usefulness of the OBO ontologies for a typical NLP task, namely coreference resolution.

3 Methods

We analyzed the OBO terms in several ways. First, we searched for OBO terms in a corpus of MEDLINE abstracts and PUBMED Central fulltext documents using different string matching strategies. Second, we investigated the overlap of OBO terms with terms from the UMLS Metathesaurus and the UMLS Specialist Lexicon. Third, we checked the OBO terms for various lexical properties and compared the distribution of these properties over the terms to that over the terms found in the two corpora and the UMLS resources. Fourth, we investigated the potential for semantic ambiguity of OBO terms. Finally, we matched the OBO terms to the coreference annotations in a pre-release of the GENIA corpus kindly provided by the Tsujii Laboratory (U Tokyo).⁵

3.1 Corpus Construction

We downloaded the OBO ontologies in May 2008 from the OBO ftp site,⁶ in OBO flat file format. 80 OBO resources were selected for investigation, excluding pure mapping or bridge files linking classes from one ontology to classes from another and files linking database entries to Go. In total, the OBO ontologies comprise 827,843 classes with different IDs. All class names and their synonyms were extracted from the 80 resources resulting in a set of 1,383,430 different OBO terms from which 791,699 are distinct class names. We included all kinds of synonyms in our study (exact, related, broad, and narrow ones) since we were interested in the complete inventory of terms provided by the OBO ontologies, rather than in exact denotations of ontology classes only.

We also downloaded the UMLS 2008AA release and used the UMLS METAMORPHOSYS tool⁷ to create a customized Metathesaurus subset

³http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

⁴http://www.nlm.nih.gov/pubs/
factsheets/umlslex.html

⁵http://www-tsujii.is.s.u-tokyo.ac.jp/

⁶ftp://ftp.fruitfly.org/pub/obo

⁷http://www.nlm.nih.gov/pubs/ factsheets/umlsmetamorph.html

OBO Ontologies			
number of ontologies	80		
distinct classes	827,843		
distinct class names	791,699		
distinct class names	1,383,430		
and synonyms			
and by nonymb			
	esources		
	esources 3,810,230		
UMLS R			
UMLS R UMLS Metathesaurus			

Table 1: Term statistics for the OBO Ontologies and for the UMLS Metathesaurus and the UMLS Specialist Lexicon

which contained all UMLS source terminologies (in English). The UMLS Metathesaurus terms (a set of 3,810,230 concept names) were extracted from the 'Concept Names File' (MRCON). Furthermore, terms from the UMLS Specialist Lexicon were extracted from the 'Agreement and Inflection' file (LRAGR) resulting in a set of 624,955 different terms. Table 1 summarizes this data for the OBO ontologies and the UMLS resources.

The MEDLINE⁸ download took place in February 2008 and included all records which contained an abstract and were entered between the years 2000 and 2008. 10% of these abstracts were randomly selected resulting in a corpus of 316,520 documents. In the following, we refer to this collection as the MEDLINE corpus.

PUBMED Central,⁹ a full-text site for biomedicine and the life sciences, was downloaded in February 2008 and, again, 10% of the documents were randomly selected. This resulted in a corpus of 6,342 documents, henceforth the PMC corpus.

The (pre-release of the) GENIA coreference corpus is composed of 1,999 documents containing in total 46,067 annotations of coreferences. Table 2 presents an overview on the number of documents and tokens contained in the three corpora we used for our study. The token counts are provided for ease of comparison only and simply rely on counting whitespace-separated strings.

Corpus	Documents	Tokens, in total
		(different tokens)
MEDLINE	316,520	65,544,220
		(1,656,388)
PMC	6,342	19,876,372
		(631,601)
GENIA	1,999	460,334 (30,004)

Table 2: Number of documents and tokens in MED-LINE, PUBMED Central (PMC) and GENIA

3.2 Matching OBO with MEDLINE and PMC

In the first part of our study, we searched for OBO terms in the MEDLINE and the PMC corpus. All OBO terms were matched against the corpora using four different strategies, *viz.* exact match, case insensitive match, case insensitive match after adding simple term variants, and case insensitive match after adding term variants and stemmed terms.

The stemmed version of a term was created using the UEA stemmer¹⁰ (for multi-token terms we only stemmed the last token of the term). The variants were generated using a combination of replacing multiple whitespaces, underscores and hyphens by blanks, and removing brackets and single quotes. In addition, we added variants for terms containing a comma followed by a space, such as "liver arginase" generated from "arginase, liver", and created some ontology-specific variants (for Go molecular function terms, e.g., which contained the suffix "activity" a variant without that suffix was added to the set of OBO terms). The matching was carried out using the LINGPIPE EXACTDICTION-ARYCHUNKER¹¹ and an integrated tokenizer.

We were interested in both, the number and nature of OBO terms that appear in the two text corpora, because these could shed light on the usage of domain-specific terms in natural language documents, and the number of OBO classes that could be identified based on these (slightly enriched) terms. Another interesting issue is whether the usage of domain-specific terminology differs in scientific abstracts (MEDLINE) from that in full-text documents (PMC). Therefore, we conducted the same matching study twice, on a corpus of abstracts and on a corpus of full-text documents.

[%]http://www.ncbi.nlm.nih.gov/sites/ entrez?db=pubmed

⁹http://www.pubmedcentral.nih.gov/

¹⁰http://fizz.cmp.uea.ac.uk/Research/
stemmer/

¹¹http://alias-i.com/lingpipe/index.html

3.3 Matching OBO with UMLS

In the next part of the study we investigated the overlap between the OBO terms and the terms provided by the UMLS Metathesaurus and the UMLS Specialist Lexicon. All OBO terms were matched against the terms from the UMLS resources ignoring case sensitivity.

The OBO ontologies hold a huge amount of domain-specific vocabulary, but hardly contain any lexical information (syntactic, morphological, and orthographic information as contained in the UMLS Specialist Lexicon) or semantic typing (contained in the UMLS Metathesaurus in terms of UMLS Semantic Network type assignments). Therefore, the overlap of OBO and UMLS terminology gives evidence of how useful the UMLS could be as a source for lexical and semantic typing information to enrich the OBO ontologies. A substantial overlap of the OBO ontologies and the UMLS Metathesaurus could reasonably be expected since five important OBO ontologies are fully or at least partially covered in the UMLS Metathesaurus as well, namely the Go, the Foundational Model of Anatomy (FMA), the NCBI taxonomy, the NCI Thesaurus, and the Medical Subject Headings (MESH).

3.4 Analysis of Lexical Properties

Inspired by the study of McCray et al. (2001), we defined a set of lexical features to analyze the properties of the OBO terms. We chose the features 'holds at least one number' (number), 'holds at least one parenthesis' (parenthesis), 'holds at least one special character' (special character) and 'average proportion of special characters' (special characters (in %)) to estimate the character complexity of the terms. The features 'average number of characters' (# characters) and 'average number of tokens' (# tokens) were selected to quantify the length of terms. We also considered the features 'holds at least one underscore' (underscore) and 'contains a comma followed by a blank' (comma space). We identified the number of OBO terms revealing these features and compared it with the number of terms in the UMLS Metathesaurus, the UMLS Specialist Lexicon, and the number of OBO terms found in the MEDLINE and PMC corpus that exhibited these features (see Table 5).

The aim of this investigation was to gather evidence, first, which types of terms appear fre-

quently in natural language documents, second, whether the nature of terms occurring in abstracts and in full-text documents differs markedly, and, third, whether the OBO terms reveal similar features as terms from the UMLS resources, or not.

3.5 Evaluating Semantic Ambiguity

Next, we analyzed the semantic ambiguity of OBO terms. We replaced all underscores in the OBO terms and turned them into lower case (case sensitivity and the use of underscores in class designators heavily depends on naming conventions) and selected exactly those terms that appeared in at least two different ontology classes and classified them as potentially ambiguous. Note that the existence of different identifiers for some ontology classes does not necessarily imply a semantic distinction (i.e., ambiguity) as well. Thus, the number of potentially ambiguous terms only constitutes an upper bound for the true number of ambiguous terms in the OBO ontologies.

The resulting list of terms was taken to determine the percentage of intra-ontology and cross-ontology ambiguity among all encountered ambiguities. We also identified the number of ambiguous terms that belonged to classes in a parent-child relationship. Our intention was to get further evidence whether the terms were really polysemous, or whether they belonged to two different ontology classes that, in fact, share the same meaning and should be merged, or whether they belonged to parent-child-related ontology classes due to a sloppy synonym assignment policy.

Semantic ambiguity in terms of polysemy (or homonymy) is a major problem when domain-specific terminology is taken into account by NLP applications (e.g., Liu et al. (2002) or Humphrey et al. (2006)). In particular, ambiguous terms decrease the performance of Named Entity Recognition (NER) tools which, in turn, affect the performance of all other NLP components relying on the output of the NER component. Thus, in order to assess the suitability of OBO terms for NLP, we consider the analysis of semantic ambiguity of the terms as an important problem.

3.6 Matching OBO with GENIA

In the last part of the study, we matched the OBO terms against the GENIA corpus, enriched with coreference annotations, and identified the number of exact and embedded matches of OBO terms

with coreference annotations (in case of an embedded match a term matches parts of a coreference annotation).

Coreferences are natural language expressions which share the same referent, i.e. refer to the same entity in the world, within or across sentences, though the denotations at the text surface are different (e.g., in "IL-7 This protein ..." "This protein" corefers to "IL-7"). The process of determining proper coreference pairs is called coreference resolution and constitutes an important subtask in many NLP applications. Expressions that are potentially coreferent can be detected by syntactic analysis (within sentences) or by some sort of discourse memory (across sentences). Domain ontologies help restrict the number of candidates for resolution by providing both, categories for semantic typing of the expressions and semantic relations between these categories that can be exploited to infer the coreference of two expressions (Vlachos et al., 2006). The aim of our study was to get preliminary evidence for the usefulness of OBO classes for semantic typing of coreferent expressions.

4 Results

4.1 Results of the Corpus Matching Study

Applying case-insensitive term matching and incorporating term variants as well as stemming (henceforth, *IVS-matching*) we found a total number of 46.7M term matches in the MEDLINE corpus and 13.2M in the PMC corpus (cf. Table 2). The matches covered major proportions of the tokens in the two corpora (76% in the MEDLINE corpus and 70% in the PMC corpus, respectively), though only minor proportions of the set of all different tokens were covered (6% in the MEDLINE corpus and 9% in the PMC corpus, respectively).

Source	OBO Classes
Ово	827,843
MEDLINE Corpus	125,386 (15%)
PMC Corpus	76,718 (9%)
UMLS Metathesaurus	528,356 (64%)
UMLS Specialist Lexicon	128,704 (16%)

Table 3: Number of OBO classes associated with the OBO terms detected in the text corpora and the UMLS resources performing IVS-matching

The main focus of the corpus matching study was on the OBO classes that can be detected in the two corpora by means of the (slightly enriched) terms associated with them. We were able to identify (see Table 3) about 125,000 OBO classes in the MEDLINE corpus (15%) and almost 77,000 in the PMC corpus (9%).

For the MEDLINE corpus we carried out additional investigations. In order to find out which impact case normalization, variant generation and stemming had on the number of traceable OBO classes we conducted additional matching experiments (exact term matching, case insensitive matching, and case insensitive matching considering term variants). By applying exact term matching only the total number of matches dropped to 25.4M and the number of OBO classes that we were able to identify decreased by four percentage points to approximatelyly 94,000. This is only 75% of the classes found by IVS-matching. Case normalization and the generation of term variants accounted for major parts of the difference, while stemming had only little impact.

Next we investigated the importance of synonyms for detecting OBO classes in the corpus. We applied IVS-matching omitting all OBO synonyms. As a result, the number of term matches dropped to 26.8M and the number of traceable OBO classes was reduced by three percentage points to approximately 97,000. This incorporated only 78% of the classes found when considering synonyms.

Finally we analyzed from which OBO ontologies the OBO classes identified in the corpus came. The study revealed that a subset of only nine (out of 80) OBO ontologies accounted for more than three-fourths of the traceable OBO classes (namely the NCI Thesaurus, the NCBI Taxonomy, the MESH, the ontology for Chemical Entities of Biological Interest (CHEBI), the Gene Ontology (GO), the INOH Molecule Role ontology (IMR), the Human Developmental Anatomy ontology (EHDA), the Foundational Model of Anatomy (FMA), and the Disease Ontology, see Table 4). The same set of ontologies holds 76% of all OBO terms and 79% of all OBO classes.

4.2 Results of the UMLS Matching Study

The focus of the UMLS matching study was on the term overlap of the OBO ontologies with UMLS

OBO Ontologies	Identified Classes in	
	MEDLINE Corpus	
NCI Thesaurus	22.26%	
NCBI Taxonomy	15.87%	
MESH	10.25%	
СнЕВІ	6.61%	
Go	6.21%	
IMR	5.61%	
EHDA	4.54%	
FMA	4.47%	
Disease Ontology	4.22%	
total	80.05%	

Table 4: Nine out of eighty OBO ontologies account for more than 80% of the OBO classes identified in the MEDLINE corpus

resources. We found (see Table 3) approximately 763,000 OBO terms that also appeared in the UMLS Metathesaurus and about 119,000 that also appeared in the UMLS Specialist Lexicon, applying IVS-matching. The OBO terms matching Metathesaurus terms were associated with about 528,000 OBO classes (64%), those matching Specialist Lexicon terms with almost 129,000 (16%).

4.3 Results of the Analysis of Lexical Properties of Terms

We found (see Table 5) the OBO terms to be on the average three tokens and 27 characters long, about twice as long compared with the OBO terms detected in the MEDLINE and the PMC corpus, and also compared with the terms in the UMLS Specialist Lexicon. Furthermore, they contained more than twice as often numbers and special characters. Compared with the terms in the UMLS Metathesaurus OBO terms were shorter and contained less often non-alphabetic characters (numers, parentheses, special characters). In addition, we found that only a small proportion of OBO terms contained underscores and 17% of UMLS Metathesaurus terms were marked with the feature 'contains a comma followed by a blank' (comma space). To locate these terms in the documents they had to be normalized first.

4.4 Results of the Evaluation of Semantic Ambiguity

We found (see Table 6) about 6% of the OBO terms to be associated with at least two ontology classes, which makes them potentially ambigu-

	Class Names	Class Names
		& Synonyms
OBO terms	1,040,119	2,013,354
ambiguous	44,193	122,491
	(4.25%)	(6.08%)
intra-source	3,816	43,747
ambiguous	(0.37%)	(2.17%)

Table 6: Number of ambiguous OBO class names and synonyms and intra-source ambiguity

ous. The average number of classes with which these ambiguous terms were associated was 2.7. Additional investigations revealed that about one third of the potentially ambiguous terms were ambiguous within one ontology and only a very small number of terms was associated with ontology classes that were related by an *is-a* relationship. When only class names were considered, i.e., synonyms were discarded, the proportion of ambiguous terms dropped by almost two percentage points and the average number of classes these terms were associated with turned out to be 2.6.

4.5 Results of the GENIA Matching Study

When OBO terms were matched with the GE-NIA corpus by IVS-matching (see Table 7) about 317,000 matches were found. These included almost 5,000 exact matches of OBO terms with coreference annotations (some 10% of all 46,000 coreference annotations) and about 40,000 matches of OBO terms that were embedded in a coreference annotation.

5 Discussion and Conclusions

We presented a study of the lexical properties of terms contained in the OBO ontologies, identified an upper bound for the semantic ambiguity of these terms and investigated how useful they are to detect references to ontology classes in domain specific text corpora.

GENIA Corpus		
number of annotated corefs	46,067	
total matches	317,493	
exactly matched corefs	4,722	
embedded matches	40,065	

Table 7: Term matches found in the GENIA corpus

Feature	OBO terms	OBO terms in	OBO terms	UMLS	UMLS
	(1,383,430)	MEDLINE	in PMC	Metathesaurus	Specialist
		(174,282)	(82,786)	terms	terms
				(3,810,230)	(624,955)
number	32%	12%	11%	34%	1%
parenthesis	10%	1%	1%	21%	0%
comma space	3%	0%	0%	17%	0%
underscore	3%	0%	0%	0%	0%
special character	24%	11%	8%	47%	16%
special characters	10	5.70	5.11	12.69	4
(in %)					
# characters	27.12	14.43	12.38	34.83	14.59
# tokens	3	1.78	1.78	4.21	1.49

Table 5: Lexical properties of terms provided by the OBO ontologies, the UMLS Metathesaurus, the UMLS Specialist Lexicon, and those OBO terms identified in the MEDLINE and the PMC corpus via IVS-matching (percentages were rounded to integers)

As far as semantic ambiguity of OBO terms is concerned we were able to characterize upper bounds assessing how many OBO terms occurred in more than one OBO class. Future work will have to constrain the grade of ambiguity by mechanisms of ontology alignment revealing semantic equivalences between ontology classes with different identifiers. After having completed the alignment of ontologies the maximal number of pair-wise non-equivalent ontology classes to which a term is assigned will reflect exactly the number of different senses of the term (excluding senses beyond the scope of the OBO ontologies). At this stage the OBO ontologies would be a valuable sense inventory for biomedical terms that could be used as a basis for word sense disambiguation (WSD). WSD is another pending step of our work. Until now we simply looked for string-based matches of OBO terms in text corpora and the UMLS, but in case the terms had different senses we did not evaluate which sense of the term turned up in the text.

We examined whether there was a difference in the use of domain-specific terminology in scientific abstracts and in full-text documents. Much to our surprise, we identified less OBO classes in full-text documents (the PMC corpus) than in abstracts (the MEDLINE corpus). Furthermore, we found OBO terms matching full-text documents to be, on the average, shorter and less complex than those matching abstracts. An explanation could be that in full-text documents OBO classes are mentioned in terms of detailed descriptions,

rather than by mentioning their name, whereas in abstracts knowledge is expressed in a much denser way, thus requiring more compact domain-specific terminology. However, the results for the PMC corpus are weaker than those for the MED-LINE corpus since it is three times smaller based on token counts. We are currently working on a second run of our experiments with a larger sized PMC corpus to reassess our results.

In the last part of our study we identified exact matches of OBO terms with GENIA coreference annotations, as well as embedded matches. What we did not do so far is checking whether in the case of two expressions matching OBO terms a semantic link between these expression could be inferred based on a semantic relation between OBO classes associated with the matched OBO terms. Only if such a relation could be found in the majority of cases coreference resolution systems would in fact benefit from the incorporation of OBO ontologies.

The main focus of our study was on the nature of OBO terms and their use in domain-specific corpora. Based on term matches we were able to infer which OBO classes were addressed in the corpora. We found that simple case normalization and term variant generation substantially increased the number of OBO classes that could be identified. Further, we observed that synonyms also played an important role.

However, the overall number of OBO classes that we were able to identify was extremely low (15% in the MEDLINE and 9% in the PMC cor-

pus). In fact, at the outset we expected a much higher proportion of OBO classes to be addressed in the domain-specific literature. Obviously, the textual realization of ontology terms is quite different from their appearance in ontologies. We assume this holds especially for those classes that are represented by rather lengthy terms containing many non-alphabetic characters. The problem of finding these additional mentions of ontology classes is well known for the Gene Ontology (cf. Blaschke et al. (2005)). Our data suggests that it also concerns the whole of OBO.

To be of real use for NLP applications we believe that the number of OBO classes that can currently be automatically detected in documents by string matching routines is still far to low. Our plans to enhance the number of traceable OBO classes are based on the following considerations. First, the OBO ontologies could be enriched with additional synonyms (presently only 39% of all OBO classes are provided with synonyms leaving room for improvement). Second, instead of simple string matching (exact or more liberal variants) more sophisticated mapping procedures should be developed particularly suited for detecting multi-token terms and terms that contain many non-alphabetic characters, such as names of chemical entities. Some specialized tools already exist that could be exploited for this purpose, such as METAMAP, a tool developed at the American National Library of Medicine to match UMLS terms and their textual forms (Aronson, 2001), or OSCAR3, a tool for the identification of chemical entities (Corbett and Copestake, 2008).

Acknowledgments

This work was partly funded by the German Ministry of Education and Research within the STEM-NET project (01DS001A-C) and by the EC within the BOOTSTREP project (FP6-028099) under the 6th Framework Program.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The METAMAP program. In Suzanne Bakken, editor, AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association, pages 17–21. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

- Christian Blaschke, Eduardo Andres Leon, Martin Krallinger, and Alfonso Valencia. 2005. Evaluation of BIOCREATIVE assessment of task 2. *BMC Bioinformatics*, 6(Supplement 1: S16).
- Peter Corbett and Ann Copestake. 2008. Cascaded classifiers for confidence-based chemical named entity recognition. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP2008)*, pages 54–62, Columbus, Ohio, USA, June.
- Susanne M. Humphrey, Willie J. Rogers, Halil Kilicoglu, Dina Demner-Fushman, and Thomas C. Rindflesch. 2006. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96–113.
- Hongfang Liu, S. B. Johnson, and Carol Friedman. 2002. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *Journal of the American Medical Informatics Association*, 9(6):621–636.
- Alexa T. McCray, Olivier Bodenreider, James D. Malley, and Allen C. Browne. 2001. Evaluating UMLS strings for natural language processing. In Suzanne Bakken, editor, *AMIA* 2001 Proceedings of the Annual Symposium of the American Medical Informatics Association, pages 448–452. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.
- Alexa T. McCray, Allen C. Browne, and Olivier Bodenreider. 2002. The lexical properties of the Gene Ontology (GO). In AMIA 2002 Proceedings of the Annual Symposium of the American Medical Informatics Association, pages 504–508. San Antonio, TX, November 9-13, 2002. Philadelphia, PA: Hanley & Belfus.
- Philip V. Ogren, Kevin B. Cohen, George K. Acquaah-Mensah, Jens Eberlein, and Lawrence Hunter. 2004. The compositional structure of GENE ON-TOLOGY terms. In *PSB 2004 Proceedings of the Pacific Symposium on Biocomputing 2004*, pages 214–225. Hawaii, USA, January 6-10, 2004. Singapore: World Scientific Publishing.
- Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. 2005. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251.
- Karin Verspoor. 2005. Towards a semantic lexicon for biological language processing. *Comparative and Functional Genomics*, 6(1-2):61–66.
- Andreas Vlachos, Caroline Gasperin, Ian Lewin, and Ted Briscow. 2006. Bootstrapping the recognication and anaphoric linking of named entities in *Drosophila* articles. In *PSB 2006 Proceedings of the Pacific Symposium on Biocomputing 2006*, pages 100–111. Hawaii, USA, January 3-7, 2006. Singapore: World Scientific Publishing.