

Semi-automated Ontology Generation for Biocuration and Semantic Search

Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

von

Dipl.-Inf. Thomas Wächter
geboren am 25. August 1978 in Jena

Betreuender Hochschullehrer: Prof. Dr. Michael Schroeder
Technische Universität Dresden

Gutachter: Prof. Dr. Lawrence Hunter
University of Colorado, Denver/Boulder, USA

Tag der Einreichung: 04.06.2010

Tag der Verteidigung: 27.10.2010

Abstract

Background: In the life sciences, the amount of literature and experimental data grows at a tremendous rate. In order to effectively access and integrate these data, biomedical ontologies – controlled, hierarchical vocabularies – are being developed.

Creating and maintaining such ontologies is a difficult, labour-intensive, manual process. Many computational methods which can support ontology construction have been proposed in the past. However, good, validated systems are largely missing.

Motivation: The biocuration community plays a central role in the development of ontologies. Any method that can support their efforts has the potential to have a huge impact in the life sciences.

Recently, a number of semantic search engines were created that make use of biomedical ontologies for document retrieval. To transfer the technology to other knowledge domains, suitable ontologies need to be created. One area where ontologies may prove particularly useful is the search for alternative methods to animal testing, an area where comprehensive search is of special interest to determine the availability or unavailability of alternative methods.

Results: The *Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG)* developed in this thesis is a system which supports the creation and extension of ontologies by semi-automatically generating terms, definitions, and parent-child relations from text in PubMed, the web, and PDF repositories. The system is seamlessly integrated into OBO-Edit and Protégé, two widely used ontology editors in the life sciences. *DOG4DAG* generates terms by identifying statistically significant noun-phrases in text. For definitions and parent-child relations it employs pattern-based web searches. Each generation step has been systematically evaluated using manually validated benchmarks. The term generation leads to high quality terms also found in manually created ontologies. Definitions can be retrieved for up to 78% of terms, child ancestor relations for up to 54%. No other validated system exists that achieves comparable results.

To improve the search for information on alternative methods to animal testing an ontology has been developed that contains 17,151 terms of which 10% were newly created and 90% were re-used from existing resources. This ontology is the core of Go3R, the first semantic search engine in this field. When a user performs a search query with Go3R, the search engine expands this request using the structure and terminology of the ontology. The machine classification employed in Go3R is capable of distinguishing documents related to alternative methods from those which are not with an F-measure of 90% on a manual benchmark. Approximately 200,000 of the 19 million documents listed in PubMed were identified as relevant, either because a specific term was contained or due to the automatic classification. The Go3R search engine is available on-line under www.Go3R.org.

Acknowledgements

The research presented in this thesis has been carried out at the Biotechnology Center (BIOTEC) of the Technische Universität Dresden.

I am extremely fortunate and grateful to have enjoyed the benefits of this excellent environment. Over the years, I received help, support, and friendship from a number of great people. It is a pleasure to thank those who made this thesis possible.

First of all, I would like to thank Michael Schroeder, my excellent supervisor, for his guidance, ideas, and support on this journey. His support was in excess of anything that I could have wished and made everything achieved possible.

Special thanks to Heiko Dietze, Andreas Doms, and Loic A. Royer for companionship, advice, trust, and professionalism right from the beginning. It is a pleasure for me to thank my colleagues Dimitra Alexopoulou, Rainer Winnenburg, and Andreas Henschel for the pleasant, effective, and successful joint work, as well as all members of the bioinformatics group for the valuable feedback and support they always provided.

I would like to express my deepest respect and gratitude to Ursula Sauer and Barbara Grune for their dedication and the confidence they showed during the creation of the Go3R search engine.

In this regard I also want to thank all former and current developers and employees of Transinsight GmbH and in particular Matthias Zschunke for his great support for this project. It has been a pleasure for me to work in such a creative and productive team.

Furthermore, I feel I must thank Dr. Monica Sturm and Prof. Michael Göttfert who during my studies gave me this insight beyond the curriculum, which made this little but so important difference.

Finally, I dedicate this work to my family. I am thankful and deeply appreciate and acknowledge their continuing support, patience, and understanding, each in their own way. Especially thank you Sylvia, thank you Elisa, and thank you Elena.

Contents

Table of Contents	iv
List of Tables	viii
List of Figures	xii
Publications	xvii
1 Introduction	1
1.1 Methods for semi-automated ontology generation	1
1.2 Automated ontology generation support for biocuration	3
1.3 Semantic search for alternative methods to animal testing	3
1.4 Overview	5
2 Background	7
2.1 Introduction	7
2.1.1 Development of biomedical ontologies	8
2.1.2 Biocuration.....	8
2.1.3 Ontology-based literature search.....	9
2.1.4 Alternative methods to animal testing.....	10
2.2 Preliminaries	11
2.2.1 Ontology.....	11
2.2.2 Biomedical ontologies and controlled vocabularies	12
2.2.3 Biomedical text-mining	18
2.2.4 Evaluation methodologies	22
2.3 Related Work: Ontology Learning	26
2.3.1 Automatic term recognition methods	27
2.3.2 Finding synonyms	37
2.3.3 Abbreviation detection	40
2.3.4 Generating textual definitions	43
2.3.5 Taxonomy generation	48
2.3.6 Availability of ontology generation methods and tools	53
2.4 Summary and Discussion	55

3	Terminology Generation	61
3.1	DOG4DAG Term Generation Method	64
3.2	Proof of concept	68
3.3	LMO Benchmark: <i>lipoprotein metabolism</i>	72
3.4	Evaluation of the quality of generated terms	72
3.4.1	Noun phrases as term candidates	73
3.4.2	Comparison of different term generation methods	75
3.4.3	Comparison of term ranking measures: C-Value vs. tf-idf	77
3.4.4	Quality of terms in dependence of the scoring method	78
3.5	Stability of the ranking	84
3.5.1	Dependency on the part-of-speech tagger	85
3.5.2	Dependency on the global corpus: Google vs. PubMed	90
3.5.3	Dependency on the ranking score: tf-idf vs. probability of occurrence	96
3.6	Summary and Discussion	101
3.7	Future Work	104
4	Definition Extraction	107
4.1	DOG4DAG Definition Extraction Method	108
4.2	Evaluation: Answering TREC2003 definitional questions	113
4.3	Evaluation: Generation of GO and MeSH definitions	116
4.4	Summary and Discussion	118
4.5	Future Work	120
5	Taxonomy Generation	121
5.1	DOG4DAG Taxonomy Generation Method	122
5.2	Evaluation: Taxonomy generation based on generated definitions	123
5.3	Pattern-based relation extraction – Superstring prediction	124
5.4	Co-occurrence analysis – Algorithm by Heymann et. al.	128
5.5	Summary and Discussion	139
5.6	Future Work	142
6	Algorithms, Data Structures and Implementations	143
6.1	TextTree – a tree representation for text	143
6.2	Taggers	144
6.2.1	AbbreviationTagger	145
6.2.2	PosTagger	145
6.2.3	NounPhraseTagger	146
6.3	Concept revisions	146
6.3.1	Merge Concept Representations	146
6.3.2	Global Frequency Revisions	148
6.3.3	Scoring Revisions	149
6.3.4	Contributions to implemented software	151

7	Integration of Ontology Generation in Ontology Editors	153
7.1	Introduction	154
7.2	OBO-Edit Ontology Generation Tool	156
7.2.1	Ontology generation in three steps.	156
7.3	Protégé Ontology Generation Plug-in	165
7.4	Go3R Ontology Editor	166
7.4.1	Design study for a new web-based ontology editor	170
7.4.2	The Ontology Editor (Version 1)	171
7.4.3	The Go3R Ontology Editor (Version 2)	172
7.5	Web-based Term Generation Platform	175
7.6	User Scenario – Biocuration	176
7.7	Summary and Discussion	180
7.7.1	Ontology learning tools	180
7.7.2	Design guidelines	181
7.7.3	Biocuration	181
7.7.4	Taxonomy editor for Go3R	182
7.7.5	Limitations	182
7.8	Future Extensions	183
7.9	Contributions	183
8	Go3R – Semantic Search for Alternative Methods to Animal Testing	185
8.1	Improving literature searches with semantic search technologies	187
8.2	GoPubMed	193
8.3	MousePubMed	196
8.3.1	Introduction	196
8.3.2	Experiment designs	200
8.3.3	Results	202
8.4	LMOPubMed	203
8.5	Go3R	205
8.5.1	Introduction	205
8.5.2	Development of the Go3R ontology	208
8.5.3	3Rs relevance filter	217
8.5.4	Recognition of Go3R ontology terms in text	220
8.5.5	Disambiguation for 3Rs methods	221
8.5.6	Author curation	226
8.5.7	Term generation for Go3R	227
8.5.8	Definition extraction for Go3R	227
8.5.9	Summary and Discussion	229
8.5.10	Future work	230
8.6	Contributions in the development of semantic search applications	231
9	Conclusion and Future Work	233
9.1	Semi-automated ontology generation	233
9.2	Automated ontology generation support for biocuration	235
9.3	Semantic search for alternative methods to animal testing	236
9.4	Future Work	237

References	I
Appendix	XIII
10.1 Related Work Summaries	XIII
10.1.1 Literature: Term Recognition Methods	XIII
10.1.2 Literature: Abbreviation detection	XVI
10.1.3 Literature: Definition Generation	XVIII
10.1.4 Literature: Taxonomy Induction	XXI
10.2 Figures and Tables	XXV
Glossary	XLIII

List of Tables

2.1	Listing of ontologies following the OBO Foundry principles	17
2.2	Listing of the main heading of the Medical Subject Headings (MeSH) .	18
2.3	The tag set of the Penn Treebank Part-of-Speech tagged corpus.	20
2.4	Contingency matrix (2x2) for two binary variable for the expected and obtained observation of the selection of some item.	22
2.5	Example data for the precision at a cutoff rank to illustrate the calculation of average precision.	24
2.6	Categorisation of term generation methods	30
2.7	Overview on term generation systems and their characteristics.	31
2.8	Overview on synonym discovery approaches regarding their characteristics and quality.	38
2.9	Pattern-matching rules for mapping an abbreviation to its full form . .	41
2.10	Overview on abbreviation detection approaches regarding their characteristics and quality.	42
2.11	Overview on the quality of definition extraction and related methods..	44
2.12	Overview on the quality of definitional question answering.	45
2.13	Distribution of how definitional statements are formulated from Westerhout and Monachesi (2008).	48
2.14	Overview on the quality of taxonomy generation.	50
3.1	Term list for “endocytosis”.	69
3.2	Top ranked terms for Hyman Group	70
3.3	Top ranked terms for Heisenberg Group	70
3.4	Top ranked terms for Buchholz Group	70
3.5	Top ranked terms for Howard Group	71
3.6	Top ranked terms for Simons Group	71
3.7	Top ranked terms for Grill Group	71
3.8	Top ranked terms for Huttner Group.	72
3.9	Queries to PubMed used to generate lipoprotein specific terminology .	73
3.10	Precision and Average Precision (rank dependent) for top 50, 200, and 1000 predictions.	75
3.11	Top 25 lipoprotein related terms generated by four methods.	76
3.12	Coverage of LMO terminology in selected document	77
3.13	Summary of best global corpus statistics at different ranks.	84

3.14	Dependency of the term generation on the choice of the Part-of-speech tagger	90
3.15	Example for changes of the term ranking in dependency of the corpus statistics.	95
4.1	Examples for Hearst patterns used for definition extraction.	113
4.2	Evaluation results for answering <i>TREC2003</i> definitional questions.	114
4.3	Example of generated answers for questions from <i>TREC2003</i> definitional question answering task.	115
4.4	Example of generated answers for questions from <i>TREC2003</i> definitional question answering task.	115
4.5	Original and the best generated definition for 4 GO and 4 MeSH terms.	117
4.6	Proportion of terms from in MeSH and GO containing parent terms, ancestor terms or other existing terms in their definitions.	118
4.7	Evaluation of generated definitions for 500 GO, 500 MeSH terms.	118
5.1	Proportion of terms from in MeSH and GO containing parent terms, ancestor terms or other existing terms in their definitions.	123
5.2	Evaluation of taxonomic information in generated definitions for GO and MeSH terms	124
5.3	Statistic on Gene Ontology terms appearance in PubMed abstracts with and without their known child terms.	125
5.4	Precision and recall observed for the top 5 and top 10 ranked potential child terms for the cases where the child terms (a) ends with and (b) starts with the parent term.	127
5.5	An extensible greedy algorithm for hierarchical taxonomy generation from social tagging systems using graph centrality in a similarity graph of tags (transcript from Heymann and Garcia-Molina (2006)).	129
5.6	Performance of the co-occurrence based generation of taxonomic relations for selected sub branches from GO and MeSH	131
5.7	Results for the reconstruction of sub-class relationships existing between 23,270 nodes in MeSH 2007.	138
6.1	Overview over software projects implemented for text mining and ontology generation.	143
6.2	Selected implementations of the class <i>ElivagarTagger</i>	145
6.3	Selected implementations of the class <i>ConceptRevision</i>	146
6.4	Statistics on tokens, sentences and n-grams in the Google Web 1T 5-gram Version 1 (Brants and Franz, 2006).	149
6.5	Statistics on the tokens and pairwise co-occurrences extracted from PubMed.	149
6.6	Examples for scoring revisions with Tf-Idf and PValue for PubMed and Google statistics	150
7.1	Listing of the source definitions for the predicted taxonomic relations for the term " <i>apolipoprotein</i> "	164

7.2	Overview over the functionalities of selected ontology editors used in the Life Sciences	167
7.3	Supported change operations of the Go3R Ontology Editor as shown in Figure 7.17.	172
7.4	Overview over the functionalities of selected ontology editors used in the Life Sciences	172
7.5	Listing for genes suggested in a Gene Ontology Annotation issues tracker requested for annotation with GO term " <i>dolichol-linked oligosaccharide biosynthetic process</i> " (GO:0006488)	177
7.6	Term candidates for PubMed query luteolysis.....	178
7.7	Extracted definitions for terms luteolysis, structural luteolysis, and functional luteolysis.	178
7.8	Listing existing terms in the Gene Ontology that end with "potassium channel activity"	179
7.9	Listing of 12 terms generated with the OBO-Edit Ontology Generation Tool for the query 'potassium channel activity'. Five terms ending with "potassium channel activity" exist in the Gene Ontology (bold), another six are good candidate terms (<i>italic</i>).	179
8.1	Overview over Gene Ontology term recognition algorithms.	189
8.2	Anatomical terms with different meanings in other knowledge domains	199
8.3	Expression patterns identified by MousePubMed in articles derived from Thut et al. (2001)	199
8.4	Overview on data sets contained in EMAGE.....	202
8.5	Quantification for facts on gene/tissue/developmental stages retrieved from literature.	202
8.6	List of the 28 branches of the Go3R ontology prototype.	212
8.7	Listing of sixteen 3Rs methods to determine eye irritation effects of substances existing in the ZEBET database.	213
8.8	5-fold cross validation for the classification of " <i>3Rs Relevant</i> " documents using Maximum Entropy model classification.	218
8.9	Definitions for selected disambiguated terms in Go3R	222
8.10	Cross validation results for disambiguation of terms in Go3R	223
8.11	Listing of the synonyms of the term 3Rs Reduction Alternative Method	223
8.12	Fifty most recent positively curated documents for " <i>3Rs Reduction Alternative</i> "	224
8.13	Fifty most recent negatively curated documents for " <i>3Rs Reduction method</i> "	225
8.14	Overview on the number of manual curations for terms in Go3R.....	226
8.15	Abbreviations extracted for animal testing alternatives related terminology	228
8.16	Lexical variants extracted for animal testing alternatives related terminology.....	228
8.17	Number of terms out of 152 terms in branch 3Rs in Human Toxicity Testing of the Go3R ontology for which a definition could be semi-automatically created.....	228

10.1	Listing of 1,000 randomly selected MeSH terms for term generation . . .	XXV
10.2	Generated terms for 1,000 MESH terms and mapping to GO, MeSH, OBO, and UMLS.	XXV
10.3	Listing of 500 randomly selected GO terms for definition generation. . .	XXV
10.4	Listing of 500 randomly selected MeSH terms for definition generation. . .	XXV
10.5	Evaluation TREC 2003: questions and manual curation of automatically generated answers	XXV
10.6	Evaluation GO definitions: listing of manual curation for top 10 generated definitions.	XXV
10.7	Evaluation MeSH definitions: listing of manual curation for top 10 generated definitions.	XXV
10.8	Proportion of terms in GO parts <i>biological_process</i> , <i>cellular_component</i> , and <i>molecular_function</i> containing parent terms, ancestor terms or other existing terms in their definitions.	XXVI
10.9	Proportion of term mentions in term definitions analysed by MeSH tree.	XXVII
10.10	Overview over correctly generated definition by <i>DOG4DAG</i> from the <i>TREC2003</i> definitional question answering task	XXVIII
10.11	Listing of 24 true negative documents which are not 3Rs relevant	XXXIII
10.12	Cross validations results for the classification of term “3Rs Relevant” with varying thresholds from 0.001 to 0.2.	XXXIV

List of Figures

2.1	Comparison Taxonomy, Thesaurus, Topic Map, and Ontology	13
2.2	Example for synonyms in the Gene Ontology	14
2.3	Examples of building principles for terms of the Gene Ontology	16
2.4	Composed Gene Ontology terms with definitional character on the example of hydrolase activity.	16
2.5	Examples for the structure of the Medical Subject Headings	17
2.6	The ontology learning cake introduced by <i>Cimiano et.al.</i>	26
2.7	WordNet entry for the word“learning”	39
3.1	The term generation pipeline	65
3.2	Illustration of the endosome biogenesis.	68
3.3	Suitability of noun phrases as ontology term candidates	74
3.4	Comparison of measures used in automatic term generation systems in the domain of animal testing alternatives: the C-Value measure vs. tf-idf for the top 500 ranked terms	78
3.5	Quality of generated terms in the lipoprotein metabolism domain	80
3.6	Pairwise comparison of the quality of term generation in dependence of the reference corpus statistics	82
3.7	Pairwise comparison of the quality of term generation in dependence of the used statistical measure.	83
3.8	Change in rank (selected samples): TNT vs. LingPipe Part-of-speech tagger	87
3.9	Change in rank (summary): TNT vs. LINGPIPE Part-of-speech tagger (part 1).....	88
3.10	Change of rank (summary): TNT vs. LINGPIPE Part-of-speech tagger (part 2).....	89
3.11	Mean precision for the retrieval of terminology in the lipoprotein metabolism domain.....	91
3.12	Summary: PubMed vs. Google based corpus statistics (part 1)	92
3.13	Summary: PubMed vs. Google based corpus statistics (part 2)	93
3.14	Selected experiments: PubMed vs. Google based corpus statistics	94
3.15	Selected experiments: probability of occurrence (PVALUE) vs. TF-IDF .	97
3.16	Summary: TFIDF vs. probability of occurrence (PVALUE) (part 1).....	98
3.17	Summary: TFIDF vs. probability of occurrence (PVALUE) (part 2).....	99

4.1	The definition recognition pipeline	109
5.1	Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for GO branches “cellular component” and “metabolic process”	132
5.2	Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for the GO branch “enzyme regulator activity” and MeSH branch “Cardiovascular System”	133
5.3	Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for the MeSH branches “Blood” and “Tissues”	134
5.4	Precision curves for centrality variants MeSH branches “Blood” and “Tissues”	135
5.5	Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for the MeSH branches “Sense Organs” and “Virus Disease”	136
5.6	Precision curves for centrality variants for the MeSH branches “Sense Organs” and “Virus Disease”	137
5.7	Generated taxonomy graph sub branch “Blood” in MeSH using co-occurrence based taxonomy generation.	140
6.1	The TextTree data structure to represent annotated text.	144
7.1	Overview on the integration of ontology learning methods in ontology editors.	154
7.2	Overview over the OBO-Edit-Ontology Generation Tool	157
7.3	OBO-Edit Ontology Tool Generation: Term generation view	159
7.4	OBO-Edit Ontology Generation Tool: Filtering by pattern in the term generation view	160
7.5	OBO-Edit Ontology Generation Tool: Filtering by existing ontology term in the term generation view	160
7.6	OBO-Edit Ontology Generation Tool: Definition generation view	162
7.7	OBO-Edit Ontology Generation Tool: reference information for generated definitions	162
7.8	OBO-Edit Ontology Generation Tool: abbreviations view	162
7.9	OBO-Edit graph viewer	163
7.10	OBO-Edit Ontology Generation Tool: Add to ontology view	164
7.11	Text editor view in OBO-Edit showing the attributes of a generated term	164
7.12	Screenshot of Ontology Generation Plugin for Protégé 4	165
7.13	Design study of the Yggdrasil Ontology Editor.	170
7.14	Screenshot of the web-based Ontology Editor (Version 1)	171
7.15	Web-based Go3R Ontology Editor (Version 2)	173
7.16	Term generation within the Go3R Ontology Editor	173

7.17	Term context menu of the Go3R Ontology Editor	174
7.18	Edit term dialog of the web-based Go3R Ontology Editor	174
7.19	The web-based Term Generation Platform has been developed for the collaborative acquisition of terminology from text. Each row shows label, abbreviations, and lexical variants found in text.	175
8.1	Semantic search in GoPubMed to answer the question which enzyme inhibits aspirin?	188
8.2	GoPubMed query for "Pax6" (continued)	194
8.3	GoPubMed query for "Pax6"	194
8.4	GoPubMed. Chart showing the absolute and relative number of publications about the disease "Aniridia" over time.	195
8.5	GoPubMed. Part of the co-authorship network for "Aniridia" in GoPubMed and author statistics showing V. van Heyningen as the author most active in this area.	195
8.6	Screenshot of MousePubMed	197
8.7	Excerpt of the anatomy ontology used in MousePubMed showing the different types of skin	198
8.8	Screenshot of LMOPubMed	204
8.9	Screenshot of Go3R	206
8.10	Go3R user interface with the search query "eye irritation" indicated in the search field	214
8.11	Refinement in Go3R using the intelligent table of contents	215
8.12	Go3R user interface, with the search query "Blood-Brain Barrier"	216
8.13	3Rs Related Categories in Go3R	217
8.14	On-line Curation Tool in Go3R	226
10.1	Listing of all 152 terms in branch 3Rs in Human Toxicity Testing of the Go3R ontology with semi-automatically created definitions.....	XXIX

Publications

Journal Papers

Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. *Bioinformatics*, 26(12):i88–i96. (ISMB 2010 Supplement), *Impact factor 2009: 4.3*

Sauer, U. G.* , Wächter, T.*, Grune, B., Doms, A., Alvers, M. R., Spielmann, H., and Schroeder, M. (2009). Go3R - semantic internet search engine for alternative methods to animal testing. *ALTEX*, 26(1):17–31, *Impact factor 2009: 1.3* *shared first author

Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., and Wächter, T. (2009). Biomedical word sense disambiguation with ontologies and meta-data: automation meets accuracy. *BMC Bioinformatics*, 10:28, *Impact factor 2009: 3.7*

Winnenburg, R., Wächter, T., Plake, C., Andreas, D., and Schroeder, M. (2008). Facts from text: Can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Briefings in Bioinformatics*, 9(6):466–478, *Impact factor 2009: 4.6*

Alexopoulou, D.* , Wächter, T.*, Pickersgill, L., Eyre, C., and Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*, 9(Suppl 9):S2, *Impact factor 2009: 3.7* *shared first author

Book Chapters

Wächter, T., Alexopoulou, D., Dietze, H., Hakenberg, J., and Schroeder, M. (2007). *Anatomy Ontologies for Bioinformatics, Principles and Practice*, volume 6, chapter Searching Biomedical Literature with Anatomy Ontologies, pages 177–194. Springer Computational Biology.

Dietze, H., Alexopoulou, D., Alvers, M. R., Barrio-Alvers, B., Doms, A., Hakenberg, J., Mönnich, J., Plake, C., Reischuk, A., Royer, L., Wächter, T., Zschunke, M., and Schroeder, M. (2008). *GoPubMed: Exploring PubMed with Ontological Background Knowledge*, chapter Bioinformatics for Systems Biology, pages 385–399, Humana Press.

Royer, L., Linse, B., Wächter, T., Bry, F., and Schroeder, M. (2006). *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, chapter Querying Semantic Web Contents: A Case Study, pages 31–52, Springer.

Doms, A., Jakoniene, V., Lambrix, P., Schroeder, M., and Wächter, T. (2006). Ontologies and text mining as a basis for a semantic web for the life sciences. In Barahona, P., Bry, F., Franconi, E., Henze, N., and Sattler, U., editors, *Reasoning Web*, volume 4126 of *Lecture Notes in Computer Science*, pages 164–183, Springer.

Conferences, Workshops, and Reports

Henschel, A., Wei Lee Woon, Wächter, T., Stuart Madnik (2009). Comparison of generality based algorithm variants for automatic taxonomy generation. In *Proceedings of 6th International Conference on Innovations in Information Technology*, December 15-17, AlAin, United Arab Emirates.

Wächter T., Tan, H., Wobst, A., Lambrix, P., and Schroeder, M. (2006). A corpus-driven approach for the design, evolution, and alignment of ontologies. In *Proceedings of Winter Simulation Conference (Computational Systems Biology)*, Monterey, CA, USA (3rd–6th December 2006), pages 1595–1602, Invited contribution.

Habich D, Wächter T., Lehner W., Pilarsky C. (2006). Two-phase clustering strategy for gene expression data sets. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC)*, Dijon, France, April 23-27, 2006, pages 145–150.

Habich D, Wächter T., Lehner W., Pilarsky C. (2006). Two-phase clustering strategy for gene expression data sets. In *LWA 2006: Lernen - Wissensentdeckung - Adaptivität*, Hildesheim, October 9th-11th 2006, pages 275–281.

Royer, L., Linse, B., Wächter, T., Bry, F., and Schroeder, M. (2006). Querying the semantic web: A case study. Forschungsbericht/research report PMS-FB-2006-31, Institute for Informatics, University of Munich.

Introduction

In the life sciences, the amount of literature and experimental data grows at a tremendous rate. The literature database PubMed alone lists nearly 20,000,000 scientific abstracts, and 700,000 are newly added every year. The protein sequence database UniProtKB stores over 10,500,000 sequences, a hundred times more than ten years ago. Turning this data into meaningful information and making it accessible to both humans and computers, is the goal of biocuration, and has become an essential part of biological discovery and biomedical research (Howe et al., 2008).

Biological and biomedical ontologies are being developed to accomplish interoperability and usability in biological databases, thus enabling efficient search and analysis. The employed ontologies are usually hierarchically structured vocabularies which are rich in synonyms and provide a taxonomy for classification. They are widely used to index and annotate data and literature in domains such as genomics, proteomics, biochemistry, animal and plant development, or anatomy (Bourne and McEntyre, 2006).

The manual construction of the required ontologies is a complex and time- and personnel-consuming effort, which involves the creation of terms, definitions, and the relations between the terms.

This thesis addresses the process of building biomedical ontologies using semi-automated ontology generation with the goal to support the needs of biocuration and ontology-based literature search in the life sciences.

1.1 Methods for semi-automated ontology generation

The Open Biological and Biomedical Ontology (OBO) Foundry, a community effort to create interoperable bio-ontologies, currently lists over 90 ontologies for various domains. All of them contain terms and taxonomic relationships, some include textual definitions.

- *Terms* are represented by single words or sequences of words (phrases) that have relevance in a given knowledge domain. Defined in an ontology, they are used for the annotation of data and for document indexing. They provide the foundation for a shared understanding of researchers who work on similar topics.
- *Definitions* describe the precise meaning of a term within the context of an ontology. They are essential as they facilitate the consistent interpretation and application of the term (Ruttenberg et al., 2007).

- *Taxonomic relationships* are the subset of the relations in an ontology that facilitates generalisation and specialisation. They form the backbone of biomedical ontologies.

Term generation is a process in which text is analysed to find exactly those words or phrases, which are suitable to represent terms as they exist in ontologies. Thereby the extracted words and phrases are ranked by relevance. Existing term extraction methods achieve adequate results by ranking long frequent multi-word terms higher in the list of candidate terms. Such methods are not applicable to short and infrequent terms. The quality of a term generation method as part of ontology generation can be judged by the method's ability to extract terms existing in a manually created ontology.

Existing approaches to definition extraction from text make use of so-called Hearst patterns, such as *A is a B*, which also indicate taxonomic relations in text. While these patterns help to find good definitions they occur infrequently in text (Hearst, 1992). An option to overcome this problem is to consider a larger document source such as the Web. The evaluation of automatic definition extraction and its overall effectiveness have recently been identified as open problems by Zhou (2007). Many of the currently available biomedical ontologies do not yet provide definitions for all terms. It has not been evaluated, if definition extraction methods are capable of finding suitable definitions for those terms.

Since definitions have the form *A is a B with property C* they provide a hint for the relation between A and B. Thus generated definitions could serve as high quality source for taxonomic relationships.

The process of creating terms, definitions, and taxonomic relationships is addressed in research question 1.

Research Question 1

To what extent can ontology construction be automated?

The goal of this work is to design and to implement ontology generation methods for the generation of terms, definitions, and taxonomic relationships in the life sciences. The methods should be fast and scalable to be suitable for integration in interactive applications. A thorough evaluation is required to build up user acceptance and to allow estimations on the overall quality of the methods.

Hypotheses:

- *Terms: Automatic term recognition methods can be extended to include single word terms while sustaining the high quality of the retrieval of domain relevant terminology.*
 - *Definitions: Pattern-based approaches to definition extraction can achieve both acceptable precision and recall when combined with suitable ranking approaches and when applied to large document sources such as the Web.*
 - *Taxonomic relationships: Definitions can serve as high quality source for taxonomic relationship extraction.*
-

1.2 Automated ontology generation support for biocuration

The manual curation of data in biological databases goes hand in hand with the extension of vocabularies and ontologies. New research results will often lead to the creation of new terms, relations, and definitions.

The two most widely used ontology editors in the life sciences are Protégé and OBO-Edit. Users manually edit their ontologies by adding term labels, definitions, and relations. To ensure the quality of such manually designed ontologies, Schober et al. (2009) recently summarised design guidelines for this process. These guidelines emphasise the use of universally understandable term labels, the inclusion of abbreviations, and the avoidance of ambiguity.

In order to contribute to the automation of ontology generation, algorithms and methods developed for research question 1 have to be integrated into the two main editors and evaluated against design guidelines as put forward in Schober et al. (2009).

Research Question 2

How can ontology generation methods be integrated into ontology editors?

The goal is to find solutions to integrate ontology generation methods into existing ontology editors used for the development of bio-ontologies. This includes the possibility to generate terms, textual definitions, and taxonomic relations, as well as the re-use of existing ontologies.

Hypothesis:

- *Ontology generation methods integrated into editors used by the biocuration community support the development of biomedical ontologies, the re-use of existing resources, and the annotation process itself.*
-

1.3 Semantic search for alternative methods to animal testing

As explained above, ontologies are widely used to make large databases and document collections accessible. One particular problem, for which ontologies may offer solutions is the search for alternative methods to animal testing.

The search for alternative methods to animal testing is not only morally and legally mandatory, but also economically advisable. In 2008, nearly 2.7 million vertebrate animals were used for scientific purposes in Germany (BMELV, 2008). This number is expected to increase further with the new EU Chemicals Regulation REACH (EC 1907/2006) that requires manufacturers to register all chemicals they use and to provide detailed information on the potential impact of each chemical on both human health and the environment. Hartung and Rovida (2009) estimated that within the EU up to 54 million vertebrate animals will be needed for testing 68,000 substances within the next 11 years. The costs of these tests have been estimated at up to 5.4 billion euro (Fauser, 2007).

Currently, the procedure of determining the availability or unavailability of alternative methods is complex and the different steps taken by the scientist are often

not transparent to others. Many potentially relevant documents are listed in literature databases, but are difficult to discover. Often, an alternative method is neither named nor is a connection to the test with animals explicitly mentioned.

Classical search technologies seek exactly what is asked for without using the context of the search terms or other information which might be relevant for the search query. Therefore, they are unable to reveal alternative methods the user has not explicitly searched for, e.g. all methods described with a synonym or more specialised term than the query term, as well as all documents found via properties of the method, rather than the methods name.

On the other hand ontology-based literature search provides an easy and comprehensive access to information stored as text (Müller et al., 2004; Doms and Schroeder, 2005; Couto et al., 2006; Dietze and Schroeder, 2008). The ontology serves as a dynamic table of contents for documents indexed with ontology terms. It also provides the context needed to distinguish between documents of general interest and documents containing specific information with relevance to alternative methods.

In the case of alternative methods to animal testing no ontology exists at the moment but many relevant aspects are modeled in existing resources. A newly developed ontology will need to organise all terminology relevant to the replacement, reduction and refinement of animal experiments, commonly referred to as the 3Rs principle of humane experimental technique (Russell and Burch, 1959). The new ontology will need to re-use general parts of existing ontologies and contain all terms needed to capture the nature of 3Rs methods as described in text. Additionally, the new ontology has to provide the labels and synonyms required to allow the identification of all occurrences of terms in text. Definitions need to be created for all terms and should establish a relationship between a term and its role in 3Rs research.

Research Question 3

How to employ ontology generation methods and ontology-based search to determine the availability or unavailability of alternative methods to animal testing.

Currently, there is no ontology for alternative methods to animal testing. How can such an ontology be created using editing tools? How applicable are automated methods for ontology generation as discussed in research question 1 and made available as answer to question 2? Finally, how can such an ontology be used to improve the search for information relevant to the 3Rs principle (Russell and Burch, 1959) and what are the limits of such an approach?

Hypotheses:

- *Ontology generation methods support the creation, extension, and maintenance of a novel 3Rs ontology.*
 - *Machine learning methods are able to relate occurrences of words to the ontology term of the correct meaning with high precision and recall. These methods can be adapted to obtain a general relevance classification for the domain of animal testing alternatives and to associate documents with specific types of methods.*
 - *Ontology-based search using a suitable 3Rs ontology allows to determine the existence of 3Rs methods in a fast, comprehensive, and transparent manner.*
-

1.4 Overview

Driven by the two application areas *Biocuration* and *Ontology-based Literature Search in the Life Sciences*, this thesis combines theoretical work on the development and validation of ontology generation algorithms with their immediate application.

The next chapter provides preliminaries and discusses the related work in the field of ontology learning. The three major contributions of the work are arranged in the following six chapters:

1. **Methods of ontology generation:** Design, development, and validation of ontology generation methods to generate terms, definitions, and taxonomic relationships
 - Development and validation of a method to generate terms for biomedical ontologies (*Chapter 3*)
 - Development and validation of a method for the extraction of definitions for biomedical ontologies (*Chapter 4*)
 - Development and validation of a method to predict taxonomic relationships (*Chapter 5*)
 - Overview of algorithms, data structures, and implementations developed for text mining and ontology generation (*Chapter 6*)
2. **Integration of ontology generation and design of ontology editors:** Design and development of the *DOG4DAG Ontology Generation Tool* integrated into the widely used ontology editors OBO-Edit and Protégé and the idea and specification of a web-based editor for the easy and fast creation of taxonomies (*Chapter 7*)
3. **Applications for ontology-based literature search:** Design and development of Go3R, the first semantic search engine for alternatives to animal testing funded by the National German Centre for Documentation and Evaluation of Alternatives to Animal Experiments (ZEBET) at the German Federal Institute for Risk Assessment (BfR) in Berlin and the Federal Ministry of Education and Research (BMBF) (*Chapter 8*)

Background

2.1 Introduction

Over the past years, numerous ontologies have been created. They are mainly used for database curation, where data, such as genes, proteins, or raw experimental data related to some species is being describe using ontology terms.

Efforts are under way to design ontologies suited not only for a single species, but rather a range of organisms. Some of these ontologies have already reached advanced stages and are widely used for annotation by many databases. One example is the Gene Ontology (GO) (Ashburner et al., 2000), a hierarchy of concepts related to biological processes, molecular functions, and cellular components used for the functional and spatial annotation of gene products. GO is one of 90 ontologies currently listed by the Open Biomedical Ontology (OBO) Foundry (Smith et al., 2007) reaching from anatomy and cell types to properties of sequences or chemical substances.

This advent of controlled vocabularies used for gene product annotation had a deep impact on life science research (Bodenreider and Stevens, 2006; Howe et al., 2008). It was a prerequisite for the analysis of high-throughput screens and cross referencing between databases of different model organisms. The use of such common ontologies that are applicable to disparate databases alleviates cross-database queries in species-centred databases like Flybase (Flybase Consortium, 1999) in the same way as in gene-centred databases like EntrezGene (Maglott et al., 2005). An example is a query across multiple species to find similarly annotated genes, possibly restricted to a common type of tissue.

Annually more than 500 publications listed in PubMed refer to the Gene Ontology even in their abstract. The calculation of enrichment in GO categories is a standard procedure in bioinformatics for the comparative analysis of genes or proteins. Just to name a few, researchers obtained functional similarity (Schlicker et al., 2007), perform data mining using Gene Ontology annotations (de Godoy et al., 2008), or employ GO for genome-wide association studies of global gene expression (Dixon et al., 2007).

Successful ontologies in the life sciences range from formal ontologies defined in description logic such as SNOMED CT via directed acyclic graphs making use of *is-a* and *part-of* relations such as the Gene Ontology to hierarchical terminologies which define narrower and broader terms like MeSH.

2.1.1 Development of biomedical ontologies

Despite the availability of ontologies in biomedical areas, specialised ontologies for different purposes are still being newly developed and maintained (Ashburner et al., 2000; Bard et al., 2005; Smith et al., 2005b; Eilbeck et al., 2005; Natale et al., 2006; Robinson et al., 2008; Mungall et al., 2010). Their creation is supported by dedicated ontology editors such as Protégé¹ and OBO-Edit (Day-Richter et al., 2007).

The development of ontologies involves many stakeholders and the costs are difficult to estimate and control. Bontas et al. (2006) conducted a survey with ontology engineers and found that it took the ontology engineers in average 5.3 month to build the ontologies with an average number of 830 ontology entities. Around 40% of the surveyed ontologies were built from scratch. For the remaining 60% of ontologies on average 50% of ontology entities have been re-used from existing ontologies. For such small ontologies with 830 entities the costs would be approximately 35,000 \$ assuming that a professional capable to model the knowledge in his or her domain costs 50,000 \$ to 90,000 \$ per year. According to this estimation the much bigger Gene Ontology would have been to build in 15 person years and would have cost millions.

This gives rise to the question how automated support for the ontology creation process can be provided. This support is needed for the development of new ontologies facilitating re-use of existing ones, as well as the extension and maintenance of existing ontologies. In the field ontology learning, researchers deal with the automated generation and maintenance of ontologies. Recently, there have been efforts to alleviate the difficulties of ontology creation and extension through text-mining which comprises a host of techniques from natural language processing to statistics. This work is reviewed in this chapter.

Overall, text-mining has to address three problems to support ontology creation and extension: (1) generation of relevant ontology terms, (2) their definitions, and (3) relationships between them.

2.1.2 Biocuration

Biocuration describes the process where scientist work to collect, annotate, and validate raw experimental data and findings from literature to make this data available in an organised form. For this biocurators employ and also participate in the development of biomedical ontologies. They facilitate communication between scientists in different communities and the interoperability between databases. Howe et al. (2008) described this role of biocurators and listed nine tasks where they work

- to extract knowledge from published papers,
- to connect information from different sources in a coherent and comprehensive way,
- to inspect and correct automatically predicted gene structures, and protein sequences to provide high-quality proteomes,
- to develop and manage structured controlled vocabularies that are crucial for data relations and the logical retrieval of large data sets,

¹ <http://protege.stanford.edu>

- to integrate knowledge bases to represent complex systems such as metabolic pathways and protein-interaction networks,
- to correct inconsistencies and errors in data representation,
- to help data users to render their research more productive in a timely manner,
- to steer the design of web-based resources, and
- to interact with researchers to facilitate direct data submissions to databases.

Many of ontologies and vocabularies listed under the umbrella of the Open Biomedical Foundry (OBO) are used for data annotation. Human annotators assign terms from such ontologies for example to genes. These assignments are ideally based on direct evidence from literature.

The goal is to make data comparable through out biological model organism databases like Flybase (Flybase Consortium, 1999), MGD (Blake et al., 2003), SGD (Cherry et al., 1998), or TAIR (Huala et al., 2001; Berardini et al., 2004). Many species-specific vocabularies have been developed covering, among others, plant (Jaiswal et al., 2005), *C. elegans* (Altun and Hall, 2006), drosophila (Grumblin and Strelets, 2006), mouse (Baldock et al., 2003; Bard et al., 1998), and human anatomy (Rosse and Mejino, 2003).

The most prominent example where biocurators use ontologies is the Gene Ontology, where scientist annotated genes and proteins with the molecular functions, biological processes, and cellular components they find in literature. Currently (as of April 11, 2010) the Gene Ontology (GO) contains 30,277 terms with 18,844 terms in “*biological_process*”, 2,727 in “*cellular_component*”, and 8,706 in “*molecular_function*”. Many widely used biological databases provide GO annotations for data on genes and proteins such as UniProt (Apweiler et al., 2004), EntrezGene (Maglott et al., 2005), and PDB (Berman et al., 2000).

Tools like GOAnnotator (Couto et al., 2006) aim to assist the curation process for GO annotations of UniProt proteins by creating the link between uncured annotations and text extracted from literature. Large model organism databases like MGI start to systematically assess text-mining systems for integration in their curation workflow (Hill et al., 2008; Dowell et al., 2009). Textpresso (Müller et al., 2004) successfully supports manual curation and recently has been estimated to speed up the curation process of *C.elegans* proteins to GO cellular components by at least 8-fold (Van Auken et al., 2009).

The proper design of exhaustive ontologies and/or controlled vocabularies to annotate, for instance, genes and gene products with structures, functions, processes, stages, or phenotypes, and their installment in relevant databases present major tasks towards facilitating comprehensive annotations and queries.

2.1.3 Ontology-based literature search

However, if terms from ontologies can be found in text, then ontologies can serve directly in literature search. Recently, a number of such knowledge-based search engines were published; for instance XplorMed (Perez-Iratxeta et al., 2003), Textpresso (Müller et al., 2004), GoPubMed (Doms and Schroeder, 2005), iHop (Hoffmann and Valencia, 2005), AliBaba (Plake et al., 2006), EBIMed (Rebholz-Schuhmann

et al., 2007), Novoseek², GoWeb (Dietze and Schroeder, 2008), and GoGene (Plake et al., 2009). To enable the creation of similar semantic application for other knowledge domains, novel ontologies are needed to be developed.

Ontology-based literature search like in GoPubMed (Doms and Schroeder, 2005) can significantly reduce search time and leads to a more comprehensive search by including search results for descendants and synonyms of the search term in the ontology. Based on these properties, ontology-based search has been identified to be of great interest for the search for information on alternative methods to animal testing.

2.1.4 Alternative methods to animal testing

In Europe, the EU Directive 86/609/EEC for the protection of laboratory animals (Commission of the European Communities, 1986) obliges scientists to consider whether any planned animal experiment can be substituted by other scientifically satisfactory methods not entailing the use of animals or entailing less animals or less animal suffering, before performing the experiment. To meet this obligation, scientists must consult the relevant scientific literature in respect to potential alternative methods prior to conducting any experimental study using laboratory animals. They need to search in text documents for information regarding the replacement, reduction and refinement of animal experiments in accordance with the *3Rs principle* (Russell and Burch, 1959). *Replacement* means that higher order animals, which are capable of suffering or feeling pain, should not be used if the aims in research, teaching, or testing can be achieved in other ways. If not avoidable, Reduction and Refinement must be applied. *Reduction* means reducing the number of animals used to obtain information of a given amount and precision, or increasing the amount of useful data obtained from the same number of animals, without compromising the quality or the quantity of animal-based research. Three main ways for reducing animal use: a) better research strategy; b) better control of variation; c) better statistical analysis. *Refinement* means any decrease in the severity of inhumane procedures applied to those animals which still have to be used.

3RS PRINCIPLE

REPLACEMENT

REDUCTION

REFINEMENT

² <http://www.novoseek.com/>

2.2 Preliminaries

Before reviewing the related work relevant ontology generation in the next section, in this section the basics used throughout the thesis are being introduced.

2.2.1 Ontology

The term *ontology* comes from the Greek word *ontologia*, and is a composition of “onto” (being) and “logia” (talking), meaning the *study of being*, or *study of existence*. The greek philosophers Platon (427 - 347BC) started to distinguish between reality and the model of the reality which defines the entities one can talk about. Later Aristotle (384 - 322BC) worked on the formalisation and the underlying logic and started to work with categories (class), *genus* (superclass / parent term), and sub-species (subclass / child term). He also introduces the notion of *differentia* to describe the difference between objects belonging to one class which allows the categorisation of the objects in different sub classes (Cimiano, 2006b). In computer science, *ontology* is used as specification of a conceptualisation (Gruber, 1993), meaning that an ontology describes the concepts of a domain and all existing relationships between them in a declarative manner. It explicitly does not describe dynamic aspects, like the transitions between states. The formal relationships in an ontology can be used to describe rules.

ONTOLOGY

GENUS
DIFFERENTIA

ONTOLOGY

In literature naming differs and is often discussed critically. In this text oriented work two most common names for these semantic units, *concept* and *class* are both used synonymously. Concept, as used here, groups a number of terms, corresponding synonyms, and abbreviations to a semantic unit, which can be referred to by all assigned terms. Concepts are defined by a natural language definition, and have a representative label (usually but not necessarily identical with one of the terms). By *term* we refer to phrases from natural language which can be simple nouns like “cell” or “growth”, or noun phrases like “early endosome”, “epidermal growth factor” which are essentially single grammatical units containing a noun as main word, here “endosome” and “factor”. More complex terms can be composed from several noun phrases like “endosomal sorting complex required for transport proteins” or “transcription factors involved in the regulation of endocytosis”.

CONCEPT
CLASS

TERM

An ontology together with a set of individual objects categorised for a class constitutes a knowledge base. The classified objects are called *instances* and represent the objects which truly exist in reality. Ontologies are widely used in biology to model the biological reality and to define the entities to allow the formulation of relations between them. Ontology stands for a rich model with high expressiveness including, terms, relations, and formal axioms. Typical models in biology, even though they are referred to as ontologies, are much simpler. In this context four levels of expressiveness can be distinguished, namely *taxonomy*, *thesaurus*, *topic map*, and *ontology*. The differences are illustrated in Figure 2.1 as well as in the following definitions.

INSTANCE

A collection of terms a community has agreed on is called a *controlled vocabulary*. Controlled vocabularies are used in biology to preserve a shared understanding between scientists. Spasić et al. (2008) defined a controlled vocabulary as a structured set of terms and definitions agreed by an authority or community. When the vocabulary is structured hierarchically one refers to it as *taxonomy*. There is only one

CONTROLLED
VOCABULARY

TAXONOMY

kind of directed relationship used to form the hierarchy. A taxonomy allows classification and simple reasoning using the directed subclass/superclass relationships which form the hierarchy. When the terms of the vocabulary are grouped by similarity is called *thesaurus*. Grouping can be achieved by the formulation of synonymy, identity, or relatedness. A thesaurus where the concepts can have properties and where all sorts of informal relations are allowed is called *topic map*. Topic maps are a tool used for knowledge representation and visualisation and has been standardised by the International Organization for Standardisation (ISO). Topic maps support the grouping of “addressable information objects around topics” and the specification of associations, the “relationships between topics”.

To represent the ontological model some representation language and format is needed. Ontologies in biology to a great extent are available in OBO format. The *OBO format* (representation language) is the text file format used by OBO-Edit, an open source, platform-independent application for viewing and editing ontologies. An OBO “[term]” entry requires a “id” and “name” to be specified. Optionally a definition can be specified as “def”, and also important synonyms as “synonym”, distinguishing between “EXACT”, “BROAD”, “NARROW”, or “RELATED”. Four built-in relationship types exist, namely “is_a”, “intersection_of”, “union_of”, “disjoint_from” for many entries, e.g. for definitions a “xrefs” can be specified as database reference.

The best known language to represent formal ontologies is OWL. The *Web Ontology Language* (OWL) is a W3C standard to formally specify ontologies to enable software systems to “understand” the meaning and relations formulated in the ontology. According to the W3C “The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics.” Several sub-languages exist with the specifications for OWL LITE, OWL DL, and OWL FULL. The different language specifications allow developers to formulate ontologies with predefined expressiveness, which has an influence on the feasibility of reasoning tasks. Reasoning is the ability to make inferences over the ontological model which lead to logically correct conclusions. Other representations of ontologies are RDF, RDFS, OIL, or F-logic which are mentioned in the related literature in this chapter, but do not play a role in the context of this work.

2.2.2 Biomedical ontologies and controlled vocabularies

A broad spectrum of biomedical terms is covered by the Gene Ontology and the Medical Subject Headings, both used in various evaluations and in the related literature the will be discussed.

Gene Ontology

The Gene Ontology Consortium develops, maintains, and uses the *Gene Ontology* (GO) (Ashburner et al., 2000), a structured, controlled vocabulary for the annotation of genes, gene products and sequences for all organisms. The GO comprises the three independent parts

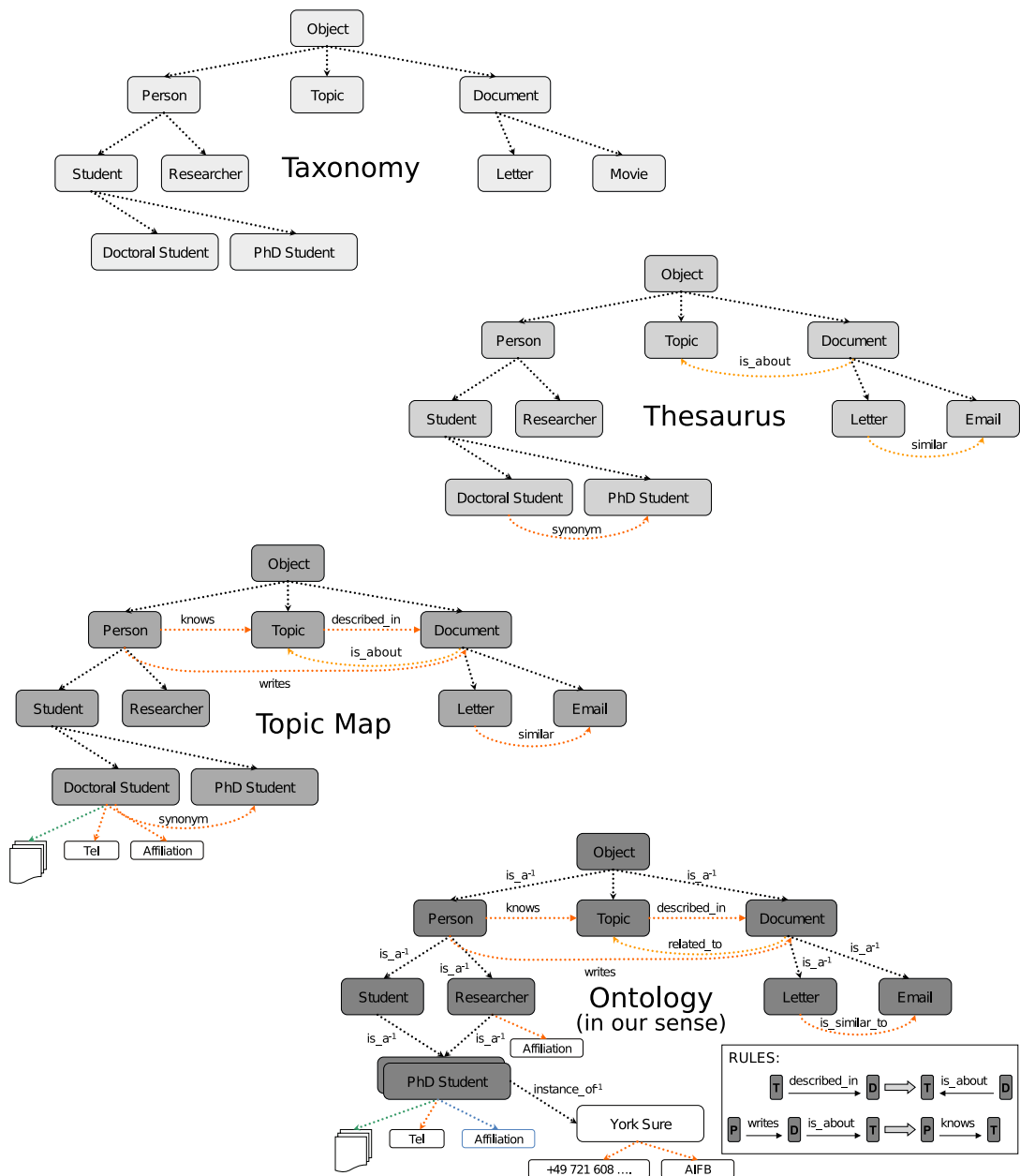


Fig. 2.1. Comparison Taxonomy, Thesaurus, Topic Map, and Ontology. An taxonomy contains hierarchically structured concepts. A thesaurus additionally allows grouping of concepts using similarity relationships. A topic map (ISO standard) allows grouping of addressable information around topics and the specification of associations between the topics. An ontology is a formal and declarative knowledge representation which aims to describes all existing concepts in a domain and the existing relationships between them. (*adapted from SemWeb 2004 tutorial by York Sure*)

- **molecular_function** for the annotation of distinct activities in the cell, like binding or transport,
- **biological_process** for the annotation more complex sequences of transition events with one or more molecular functions involved, and
- **cellular_component** for the annotation of locations of action within the cell.

The GO contains ontological defined relations, *is_a* and *part_of* between terms. There does not exist relationships between the three parts. Relations state relationships between classes not instances. In most cases the relations are stated as universal quantification (\forall , “for all”), rather than as existential quantification (\exists , “for some”). Graph-theoretically the GO is directed-acyclic graph, meaning that no cycles exist and that all relations used are directed. Currently (as of April 11, 2010) the GO contains 30,277 terms with 18,844 terms in “*biological_process*”, 2,727 in “*cellular_component*”, and 8,706 in “*molecular_function*”. More than 99.2% of GO terms have a definition. The GO guidelines suggest the used of generic term definitions. A term should be defined through the parent term by describing the difference (differentia) of instances of the specific term to instances of other sibling terms. For each definition it is intended to have a reference to the creator and/or source of the information.

Each GO term can have exact, broader, or narrower synonyms. For example, the term “*mitochondrial chromosome*” (GO:0000262) has the narrow synonym “mitochondrial DNA”, because the chromosome consist of (has_part) DNA and DNA does not correspond to the whole chromosome. “Mitochondrial genome” is a related synonym as genome might refer to one, but also to a set of chromosomes depending on the species.

```
[Term]
id: GO:0000262
name: mitochondrial chromosome
namespace: cellular_component
def: "A chromosome found in the mitochondrion of a eukaryotic cell." [GOC:mah]
synonym: "mitochondrial DNA" NARROW []
synonym: "mitochondrial genome" RELATED []
synonym: "mtDNA" NARROW []
is_a: GO:0000229 ! cytoplasmic chromosome
is_a: GO:0044429 ! mitochondrial part
relationship: part_of GO:0042645 ! mitochondrial nucleoid
```

Fig. 2.2. Example for synonyms in the Gene Ontology. Term definition in OBO 1.2 format for the GO term “*mitochondrial chromosome*” (GO:0000262) with the narrow synonyms “mitochondrial DNA” and the related synonym “mitochondrial genome”.

The term “*transcription export complex*” [GO:0000346] has the exact synonym “TREX complex” in this case an abbreviation, and “*transport vesicle*” [GO:0030133] has a broader synonym “secretory vesicle”, because transport vesicle are capable to secrete molecules they previously internalised, but they additionally can transport these molecules throughout the cell.

The GO term labels are unique and follow certain syntactical creation patterns. Intrinsic building principles re-occur throughout the ontology. Ogren et al. (2004) analysed the structure of the GO and highlighted that the semantic relationships also coincide with clear surface linguistic relationships and gives as example: mem-

brane [GO:0016020] *has_part* inner membrane [GO:0019866] *has_part* mitochondrial inner membrane [GO:0005743] *has_part* mitochondrial inner membrane peptidase complex [GO:0042720]. Another examples for term labels containing the parent term label is “*nerve-nerve synaptic transmission*” contain the parent “*synaptic transmission*” (Figure 2.3). Ogren et al. (2004) showed that 65.3% of labels and synonyms in the GO at the time contained a label or synonym from another term. This affected in total to 72% of GO terms, which include other GO terms. For the regulation of biological processes terms frequently have a prefix “negative” (15%) and “positive”(15%) to qualify the direction of regulation like for the term “*negative regulation of action potential*”. To describe complex biological functions and processes, GO terms are often composed from several components giving a term some definitional character like “*hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides*”. The compositional structure of GO terms has been studied by Yeh et al. (2003) and Mungall (2004). The composition of terms lead to redundancy in textual definitions and relationships, but serves the purpose of biocuration. In many cases only one term need to be annotated to describe a complex biological process or molecular function. The different structural units of composed terms could also be modelled separately as shown in Figure 2.4 for “*hydrolase activity*”. Due to the terms complexity they have a low probability to literally occur in text, making the automatic annotation of text with GO terms a difficult problem (Doms, 2009). In general it can be said, that the deeper terms are in the hierarchy the longer the labels tend to be.

True-path rule The Editorial Style Guide for the development of the GO describes the so called “true path rule” which states, that “the pathway from a child term all the way up to its top-level parent(s) must always be true”. Following this, it is assured, that a biological entity annotated with a term, implies biologically correct the annotation with any ancestor of this term. This property allows comparability of annotated biological data at different levels, but also enables semantic search as used in GoPubMed (Doms and Schroeder, 2005).

Open Biological and Biomedical Ontologies (OBO)

The GO is part, maybe the core, of the OBO ontologies. The *Open Biomedical Ontologies (OBO) Foundry*, a community effort to create interoperable bio-ontologies, lists currently over 90 them. The OBO lists such ontologies, that adhere to the OBO Foundry principles and are (1) freely available, (2) use a common shared syntax, (3) have a unique identifier space, are (4) versioned, (5) have clearly specified content and purpose, (6) provide definitions for all terms, (7) use the relations of the OBO Relation Ontology, (8) are well documented, (9) are used by several independent users, and (10) are developed collaborative. The parts of Gene Ontology as well as other ontologies have this status (Table 2.1), all others are listed as candidate ontologies.

OBO FOUNDRY

Medical Subject Headings (MeSH)

The use of structured vocabularies by libraries for document retrieval has a long history. Categorised lists of terms were printed for the first time in the 1963 Medical Subject Headings of U.S. National Library of Medicine (NLM) and contained

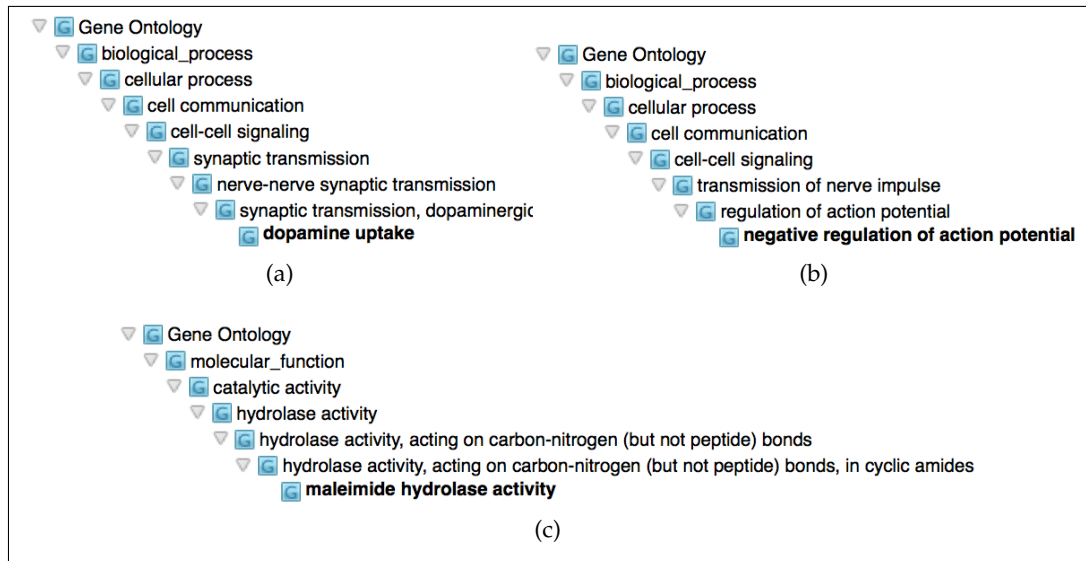


Fig. 2.3. Examples of building principles for terms of the Gene Ontology. (a) “Dopamine uptake” has as ancestors “synaptic transmission” with child “nerve-nerve synaptic transmission” which contains the parent term. (b) Regulation terms often contain the prefix “negative” or “positive” like “negative regulation of action potential” which is child of “regulation of action potential”. (c) GO terms can be composed from other terms like “hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides”.

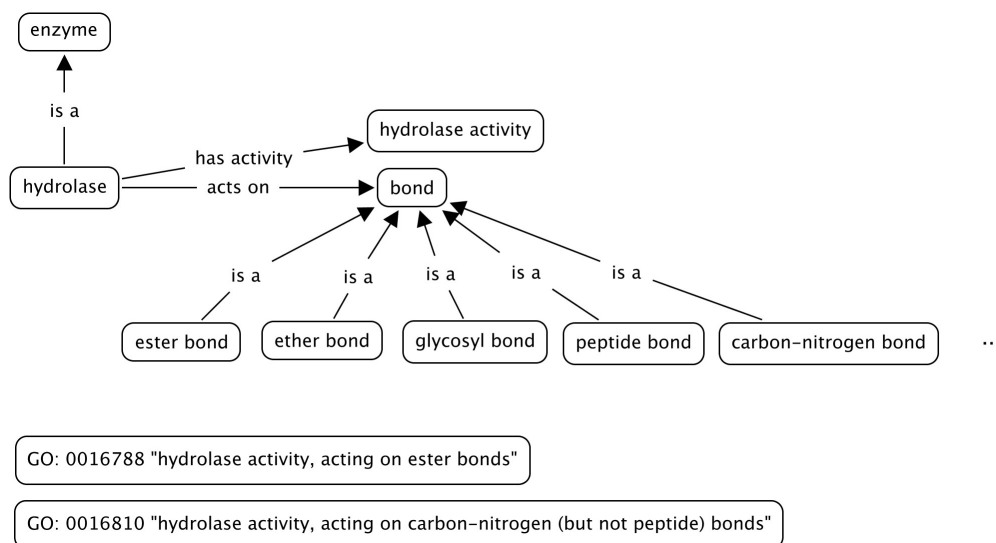


Fig. 2.4. Composed Gene Ontology terms with definitional character on the example of hydrolase activity. Terms like hydrolase, hydrolase activity, bond, ester bond and relations between them (e.g. acts on) can be easily found in text, whereas full GO terms such as “hydrolase activity, acting on ester bonds” are unlikely to appear literally in an article.

Title	Domain	Prefix
Biological process	biological process	GO
Cellular component	anatomy	GO
Chemical entities of biological interest	biochemistry	CHEBI
Molecular function	biological function	GO
Phenotypic quality	phenotype	PATO
PRotein Ontology	proteins	PRO
Xenopus anatomy and development	anatomy	XAO
Zebrafish anatomy and development	anatomy	ZFA

Table 2.1. Listing of ontologies following the OBO Foundry principles. 8 of over 90 ontologies have currently the status of OBO Foundry ontologies. This means they adhere to the principles and are (1) freely available, (2) use a common shared syntax, (3) have a unique identifier space, are (4) versioned, (5) have clearly specified content and purpose, (6) provide definitions for all terms, (7) use the relations of the OBO Relation Ontology, (8) are well documented, (9) are used by several independent users, and (10) are developed collaboratively.

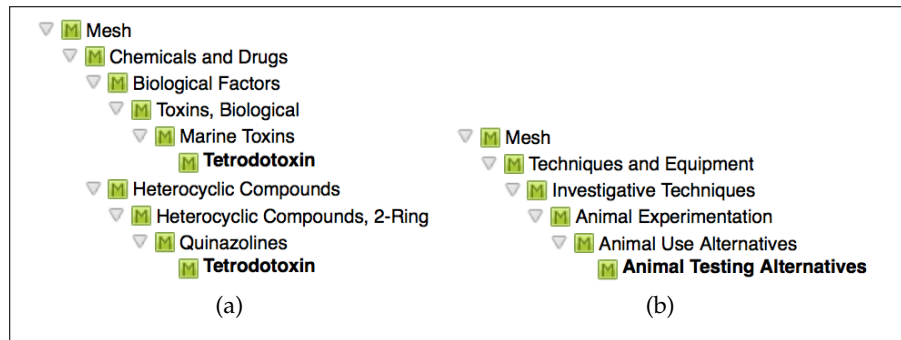


Fig. 2.5. Examples for the structure of the Medical Subject Headings MeSH is a thesaurus used for document retrieval, where terms are grouped by topic. (a) Terms can be categorised in several topics, here “*Tetradotoxin*” as marine toxin and as heterocyclic compound. (b) “*Animal Testing Alternative*” is therefore descendant of “*Animal Experimentation*”, which is the topic classification not a subsumption relation.

thirteen main categories and a total of fifty-eight separate groups in subcategories and main categories. Today, the Medical Subject Headings (*MeSH*) are a thesaurus with 25,588 descriptors arranged in 11 high level branches, such as e.g. Anatomy or Diseases & Symptoms (Table 2.2), and is used to index the 20 million scientific literature abstracts of *MEDLINE* which come from 5,400 biomedical journal. The data can be conveniently accessed via *PubMed*, a service developed and maintained by the National Center for Biotechnology Information (NCBI) at the NLM. PubMed also provides access to fulltext article listed on PubMed Central. The MeSH have a hierarchical structure that allows searches at various levels of specificity (Figure 2.5). In the MeSH 2010, the 25,588 descriptors are enriched with 172,000 English Entry Terms (some are synonyms) to give evidence for an appropriate association of bibliometric references to MeSH headings. Such entry terms exist for several languages, including German. The MeSH headings itself are translated in several languages, including Czech, Dutch, French, German, Italian, Japanese, Portuguese, Spanish, or Swedish.

MESH

MEDLINE

PUBMED

Main heading	Sub tree
Anatomy	[A]
Organisms	[B]
Diseases	[C]
Chemicals and Drugs	[D]
Analytical, Diagnostic and Therapeutic Techniques and Equipment	[E]
Psychiatry and Psychology	[F]
Phenomena and Processes	[G]
Disciplines and Occupations	[H]
Anthropology, Education, Sociology and Social Phenomena	[I]
Technology, Industry, Agriculture	[J]
Humanities	[K]
Information Science	[L]
Named Groups	[M]
Health Care	[N]
Publication Characteristics	[V]
Geographicals	[Z]

Table 2.2. Listing of the main heading of the Medical Subject Headings (MeSH). The top level of MeSH is provided by 16 main headings which structure the 25,588 MeSH descriptors.

Unified Medical Language System (UMLS)

With the aim to reduce the fundamental barriers to the application of computers to medicine, the NLM assembled a large multidisciplinary, multisite team to work on the *Unified Medical Language System (UMLS)* in 1986. The UMLS consists of three major sources, the UMLS Metathesaurus, the UMLS Semantic Network, and the UMLS SPECIALIST Lexicon and has the purpose to facilitate the development of computer systems with awareness of biomedical vocabulary and relations between them. The Metathesaurus integrates many different vocabularies, thesauri, and ontologies and provides a single database format to access all. MeSH, GO, and many OBO ontologies are part of the UMLS Metathesaurus. The UMLS Semantic Network adds the structure between Metathesaurus concepts, primarily via *is_a* links between concepts. Non-hierarchical relationships are used to formulate relatedness, namely ‘physically related to’, ‘spatially related to’, ‘temporally related to’, ‘functionally related to’, and ‘conceptually related to’. The SPECIALIST lexicon intends to be a general English lexicon that includes many biomedical terms. Its purpose is to support Natural Language Processing and provides syntactic, morphological, and orthographic information. The UMLS is a great resource for the benefit to biomedical text-mining.

2.2.3 Biomedical text-mining

In biomedical text mining, researchers use lexical, syntactic, and semantic techniques to extract desired information from text (Jensen et al., 2006). Related research fields are Natural Language Processing and computational linguistics, as well information retrieval including machine learning and word sense disambiguation. *Natural Language Processing (NLP)* is an area of computer science which deals with the interaction of computers with humans by parsing, interpreting, or generating human

natural languages by using techniques provided by computational linguistics such as statistical or rule-based modeling of natural languages.

Before text is analysed or interpreted a number of standard process steps are usually performed. An input stream of characters of a text is tokenized, meaning that *tokens* are obtained. The tokens are being categorised in pre-defined classes standing for types of symbols, like number, punctuation, comma, opening or closing bracket, words, etc. These classes help to specify algorithms to process the text. *Sentence splitting* assembles the tokenized text into sentences. A difficulty in sentence splitting is to decide whether a possible punctuation mark is a true delimiter or is part of some textual unit within a sentence like a organism (“C. elegance”) or a person name (“Mr. Smith”). For ontology learning obtaining the structural units of the text is of greater interest, meaning sentence splitting, the identification of tokens and noun phrases, as well as the awareness for term variations and normalisation using stemming and dictionaries.

TOKEN

SENTENCE
SPLITTING

- *morphological*: inflection e.g. singular vs. plural;
- *orthographic*: hyphens, slashes, upper case, lower case, etc
- *lexical*: lexical synonyms e.g. “cancer” vs. “carcinoma” ;
- *structural*: use of prepositions e.g. “clones of human” vs. “human clones”;
- *acronyms and abbreviations*:

Stemming is capable of resolving morphological variations by obtaining a normalised base form of each term (Porter, 1997). Stemming is fast and simple, but introduces additional ambiguity. Very often, words will appear in different forms, such as “binding” and “binds”. These refer to the same concept, which can be solved by resolving words to their stem (“bind”). However, the analogous reduction of “dimerisation” to “dimer” is more questionable. The former talks about the process, the latter about the result. A similar example is “organisation”, where a transformation into “organ” is invalid as well as “sensitive” and “sensitisation”, both stemmed to “sensiti” suggesting equality but are in fact different, as one is a property while the other one describes a whole process.

STEMMING

Part-of-speech tagging

With *Part-Of-Speech (POS)* the grammatical classification, or the syntactic category a word is denoted, to which a word can be assigned to in the context of a phrase, sentence or paragraph. This categories can be many fold and can be mapped to classes like noun, adjective, adverb, verbal. POS tagging is the next step of making use of linguistic knowledge to interpret the tokens obtained from text. The concrete categories depend on the annotated categories in the annotated corpus. As example the tags used in the Penn Treebank corpus (Marcus et al., 1993) are listed in Table 2.3. An example sentence has been tagged for Example 2.1.

POS

Noun phrase chunking

Phrase chunking divides sentences into non-overlapping sequences of tokens. Noun phrase chunking recognises chunks that consist of noun phrases (NP). Other tasks are recognising verbal phrases, pronoun phrases, or participle phrases. For term

PHRASE
CHUNKING

CC	Coordinating conjunction	TO	to
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present participle
IN	Preposition/subord. conjunction	VCN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sing. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sing. present
JJS	Adjective, superlative	WDT	wh-determiner
LS	List item marker	WP	wh-pronoun
MD	Modal	WP\$	Possessive wh-pronoun
NN	Noun, singular or mass	WRB	wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	\$	Dollar sign
NNPS	Proper noun, plural	.	Sentence- nal punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(Left bracket character
PP\$	Possessive pronoun)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	'	Left open single quote
RBS	Adverb, superlative	"	Left open double quote
RP	Particle	'	Right close single quote
SYM	Symbol (mathematical or scienti c)	"	Right close double quote

Table 2.3. The tag set of the Penn Treebank Part-of-Speech tagged corpus.

recognition (Section 2.3.1), the notion of a noun phrase as term candidate is of interest. A *noun phrase* is a sequence of words, that are a unit and can act as subject, complement, or object in a sentence. A recent overview by Wermter et al. (2005) evaluated the performance of state-of-the-art machine learning based noun phrase chunkers for biomedical text. The chunkers have been trained on the PENN TREE-BANK newspaper corpus and tested on the biomedical text corpus (GENIA). The results on GENIA have been 3-6% lower depending on the system. In Example 2.1 the noun phrases are shown in square brackets extracted after POS tagging.

NOUN PHRASE

Example 2.1 (Part-of-Speech tagging with the Stanford POS-tagger). A sentence from a PubMed abstract (PMID 19442486) was Part-of-Speech tagged using the Stanford POS-tagger (Toutanova and Manning, 2000) and the noun phrases have been extracted with the noun phrase chunker by Ramshaw and Marcus (1995).

Sentence:

The mouse embryonic stem cell test (EST) was designed to predict embryotoxicity based on the inhibition of the differentiation of embryonic stem cells (ESC) into beating cardiomyocytes in combination with cytotoxicity data in monolayer ESC cultures and 3T3 cells.

POS-tagged sentence:

The/DT mouse/NN embryonic/JJ stem/NN cell/NN test/NN -LRB-/-LRB- EST/NNP -RRB-/-RRB- was/VBD designed/VBN to/TO predict/VB embryotoxicity/RB based/VBN on/IN the/DT inhibition/NN of/IN the/DT differentiation/NN of/IN embryonic/JJ stem/NN cells/NNS -LRB-/-LRB- ESC/NNP -RRB-/-RRB- into/IN beating/VBG cardiomyocytes/NNS in/IN combination/NN with/IN cytotoxicity/JJ data/NNS in/IN monolayer/NN ESC/NN cultures/NNS and/CC 3T3/CD cells/NNS ./.

NP chunking (NPs in square brackets):

[The/DT mouse/NN embryonic/JJ stem/NN cell/NN test/NN -LRB-/-LRB- EST/NNP -RRB-/-RRB-] was/VBD designed/VBN to/TO predict/VB embryotoxicity/RB based/VBN on/IN [the/DT inhibition/NN] of/IN [the/DT differentiation/NN] of/IN [embryonic/JJ stem/NN cells/NNS -LRB-/-LRB- ESC/NNP -RRB-/-RRB-] into/IN beating/VBG [cardiomyocytes/NNS] in/IN [combination/NN] with/IN [cytotoxicity/JJ data/NNS] in/IN [monolayer/NN ESC/NN cultures/NNS] and/CC [3T3/CD cells/NNS] ./.

Word Sense Disambiguation

Word sense disambiguation (WSD) is a sub-task of semantic tagging and deals with relating the occurrence of a word in a text to a specific meaning, which is distinguishable from other meanings that can potentially be related to that same word (Schuemie et al., 2005). WSD is essentially a classification problem: given an input text and a set of sense tags for the ambiguous words in the text, assign the correct senses to these words. Sense assignment often involves two assumptions: a. within a discourse, e.g. a document, a word is only used in one sense (Gale et al., 1992) and b. words have a tendency to exhibit only one sense in a given collocation – neighbouring words (Yarowsky, 1993). (Alexopoulou et al., 2009) analysed and evaluated 4 approaches to word sense disambiguation. The ‘Closest Sense’ method assumes that the ontology defines multiple senses of the term. It computes the shortest path of co-occurring terms in the document to one of these senses. The ‘Term Cooc’ method defines a log-odds ratio for co-occurring terms including co-occurrences inferred from the ontology structure. The ‘MetaData’ approach (Doms, 2009, chapter: Algorithms for Concept Recognition) trains a maximum entropy classifier on meta-data, such as journal, author, date of publication. It does not require any ontology, but requires training data, which the other methods do not. To evaluate these approaches we defined a manually curated training corpus of 2,600 documents for seven ambiguous terms from the Gene Ontology and MeSH. All approaches over all conditions achieve 80% success rate on average. The ‘MetaData’ approach performed best with 96%, when trained on high-quality data. Its performance deteriorates as quality of the training data decreases. The ‘Term Cooc’ approach performs better on Gene Ontology (92% success) than on MeSH (73% success) as MeSH is not a strict

is-a/part-of, but rather a loose is-related-to hierarchy. The 'Closest Sense' approach achieves on average 80% success rate. Alexopoulou et al. concluded that metadata, such as journal, author name is valuable for disambiguation, but requires high quality training data. The closest sense method requires no training, but a large, consistently modelled ontology, which are two opposing conditions. Term co-occurrence achieves greater 90% success given a consistently modelled ontology. Overall, the results show that well structured ontologies can play a very important role to improve disambiguation.

MAXIMUM
ENTROPY

Maximum entropy method The *maximum entropy* method was introduced by Berger et al. (1996) and is a method for statistical modelling where minimal assumptions are made about the data. The method allows the assignment of a-priori probability to known classes based on incomplete information. As the name suggest, the method aims to maximize the entropy and the authors describe the methods goal as follows: *model all that is known and assume nothing about that which is unknown. In other words, given a collection of facts, choose a model consistent with all the facts, but otherwise as uniform as possible.* In information theory, *entropy* (or self-information) measures the amount of information in associated with a random variable (Manning and Schütze, 1999).

ENTROPY

$$HX = - \sum p(x) \log_2 p(x),$$

with $p(x)$ the probability mass function of a random variable X , over a discrete set of symbols.

2.2.4 Evaluation methodologies

While the final judgment on the quality of a system should be based on the application task, Natural Language Processing systems are usually evaluated and compared using standardised technical measures such as precision, recall, and F-measure. The basis for this measure are the sets resulting from the overlap of the obtained selection of items and the expected selection of items, which are the correct selected expected items (*true positives*), the wrongly selected items (*false positives*), the expected items the method missed to select (*false negatives*), and the items not selected and not expected to be selected (*true negatives*).

TRUE POSITIVE
FALSE POSITIVE
FALSE NEGATIVE
TRUE NEGATIVE

obtained	expected	
	1	0
1	true positive	false positive
0	false negative	true negative

Table 2.4. Contingency matrix (2x2) for two binary variable for the expected and obtained observation of the selection of some item.

PRECISION

Definition 2.2 (Precision). The measure precision (or specificity) is defined as the proportion of selected items that a method selected correctly.

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

A high precision indicates that most retrieve items have been correct. A low precision means that a system retrieved many incorrect items. A precision of 1 means that a system retrieves only correct item.

Definition 2.3 (Recall). The measure recall (or sensitivity) is defined as the proportion of items a method selected. RECALL

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

A high recall indicates that most of what could have been found was found. A low recall means that a system failed to find what should have been found. A recall of 1 can theoretically be achieved by just returning everything. Therefore, for most evaluation tasks the performance of a system is a trade-off between precision and recall. For an easier comparison of systems Van Rijsbergen (1979) introduced the F-measure, as the harmonic mean between precision and recall.

Definition 2.4 (F-measure). The F-measure (F) (or F-score) is defined as the harmonic mean between precision and recall. A value $\alpha = 0.5$ is equal weighting recall and precision. F-MEASURE

$$F - measure = \frac{1}{\alpha \frac{1}{precision} + (1 - \alpha) \frac{1}{recall}}$$

Learning accuracy Sometime Natural Language Processing systems are also evaluated in terms of *learning accuracy*. The measure was introduced by Hahn and Schnattinger (1998). It “measures not only the overall correctness of the final classification but also incorporates the distance between the position f predicted by the algorithm and the correct one s ”, see Witschel (2005). LEARNING ACCURACY

Accuracy Generally *accuracy* is defined as the percentage of items selected correctly (true positives + true negatives) and the corresponding error is defined as the percentage of wrongly selected items (false positives + false negatives). ACCURACY

Average precision When measuring the performance of rankings of elements (e.g document or term rankings) also the recall at a certain rank needs to be reflected in the measure. The measurement of precision only takes the correct or relevant elements in the set of retrieved elements into account. *Average precision* does incorporate recall at a rank and is therefore a retrieval order dependent precision measure. Overall the average precision values are smaller or equal then precision values depending on the number of not relevant elements retrieved prior the retrieval of all relevant elements. Average precision can be defined as follows: AVERAGE PRECISION

Definition 2.5 (average precision). The average value of the precision p at rank r , $p(r)$ is defined as

$$avgP = \frac{\sum_1^N p(r) * rel(r)}{Number\ of\ relevant\ elements\ retrieved}$$

where $rel(r)$ is a binary relation returning 1 if the element at rank r is relevant and 0 otherwise. and $p(r)$ is the precision at a cutoff rank r .

To illustrate the underlying principle see the example below of how to calculate average precision for binary and probabilistic values.

Example 2.6. In the following table there are two example shown with overall the same precision of 1.0 but different rankings.

rank(i)	Example A			Example B		
	rel(i)	p(i)	r(i)	rel(i)	p(i)	r(i)
1	1	$\frac{1}{1}$	$\frac{1}{3}$	1	$\frac{1}{1}$	$\frac{1}{3}$
2	1	$\frac{2}{2}$	$\frac{1}{3}$	1	$\frac{2}{2}$	$\frac{1}{3}$
3	0	$\frac{2}{3}$	0	1	$\frac{3}{3}$	$\frac{1}{3}$
4	0	$\frac{2}{4}$	0	0	$\frac{3}{4}$	0
5	1	$\frac{3}{5}$	$\frac{1}{3}$	0	$\frac{3}{5}$	0
6	0	$\frac{3}{6}$	0	0	$\frac{3}{6}$	0

Table 2.5. Example data for the precision at a cutoff rank to illustrate the calculation of average precision.

The average precision $avgP(A)$ and $avgP(B)$ can be calculated as follows.

$$avgP(A) = \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5} \right) * \frac{1}{3} = 0.87$$

$$avgP(B) = \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{3} \right) * \frac{1}{3} = 1.0$$

No matter which measure was chosen for evaluation it has to be judged on the meaningfulness of measure on a case-by-case basis. Several difficulties have been discussed in literature. For text-mining dependent systems an evaluation can either be done based on an experts judgement or based on a gold standard. A *gold standard* can be a data set produced by a method that is widely accepted as being the best available or it can be manually created by experts to compare different systems. Gold standard evaluation are common in biomedical information retrieval but few benchmarks are available. Known reference corpora (solutions) for text retrieval are created at the Text REtrieval Conference (TREC) workshops which already had tracks on the retrieval of genomic data, general question answering, recently on large scale search in chemistry-related documents, and explore information seeking behaviors common in general web search. For the task of named entity recognition (NER) Hakenberg (2007) named five facts, that make the evaluation difficult. Four of them are general to evaluation task in biomedical text mining.

1. **Availability of corpora (data sets):** Few corpora are available, that are sufficiently large for meaningful comparisons. Very often tools are only evaluated on 10 - 100 PubMed abstracts.
2. **Annotation is subjective:** for NER in particular it cannot be assumed that the annotator is aware of all gene and protein names. In other areas, maybe not

all annotators will have to have wider understanding to decide on the validity of an annotation. Therefore annotation guidelines are important and the same document or data should be annotated by several persons to obtain the inter-annotator agreement.

3. **Matching accuracy:** One can be variably strict in the decision on true positives. For question answering one could require for instance the only exact answer, allow similar but correct answers, or reward correct facts contained in answers. Not quite correct answers can be penalised twice as they are counted as false negative for the missed fact and as false positive for the not quite correct retrieval, and other way around.
4. **Unbiased evaluation:** The evaluation should always be performed in a task-specific manner to avoid and recognise the tuning of methods for just one of the tasks.

2.3 Related Work: Ontology Learning

The ontology learning process (Figure 2.6) consists of several sub-tasks, namely the discovery of terms, concepts, synonyms/abbreviations, taxonomy, relations and rules/axioms (see Cimiano (2006a,b)). The task *Terms* corresponds to the actual observed phrases in text. *Synonyms* denotes the grouping of terms with same or similar meaning. The task *Concepts* deals with the creation of formal representations defining the Intension (Int) – an informal definition, Extension (Ext) – the instances described by the definition, and Lexicon (Lex) – the term and its synonyms. Subsumption relations are learned in task *Taxonomy*, while other specific associations are learned in task *Relations*. Finally, actual facts about the concepts are collected in the task *Rules&Axioms*.

This section covers the relevant literature associated with the ontology learning sub tasks. For the task *Concepts* the associated topics *finding synonyms*, *abbreviations detection*, and *finding textual definitions* have been reviewed. For these and for the retrieval of *terms* and *taxonomy* an overview over the relevant literature is given followed by a summary including an estimation in the methods applicability and availability. This section on related work is organized in the following sections:

- Term recognition methods (Section 2.3.1)
- Finding synonyms (Section 2.3.2)
- Abbreviations detection (Section 2.3.3)
- Finding textual definitions (Section 2.3.4)
- Taxonomy generation (Section 2.3.5)

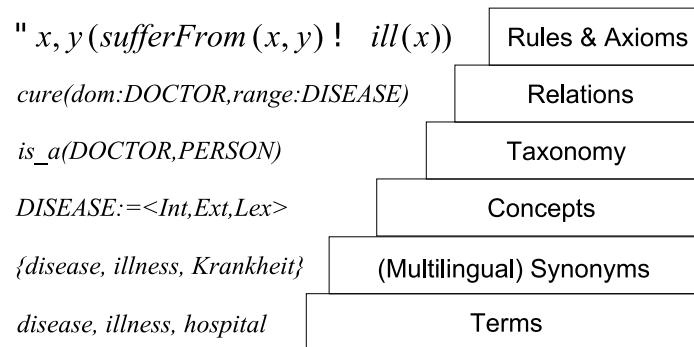


Fig. 2.6. The ontology learning cake introduced by Cimiano (2006a). *Terms* correspond to the actual observed phrases in text; *Synonyms* denote a group of terms with same or similar meaning; *Concepts* are formal representations defining the Intension (Int) – an informal definition, Extension (Ext) – the instances described by the definition, and Lexicon (Lex) – the term and its synonyms; *Taxonomy* means the learning of subsumption relations, while *Relations* deals with other specific associations; and *Rules&Axioms* state actual facts about the concepts using the relationships.

2.3.1 Automatic term recognition methods

Automatic term recognition (ATR) is the extraction of domain relevant terms from natural language text using linguistic and statistical information. ATR helps to extract the terms in the first ontology learning sub task. The extraction of terms from text is a two step procedure:

ATR

- **Step 1:** Extraction of the terms by finding begin and end of each candidate term
- **Step 2:** Filtering of term candidates to reduce the terms from step 1 to the relevant terms. This also involves grouping and the classification of terms.

Table 2.7 summarises 16 existing term recognition approaches, including early methods reviewed by Castellví et al. (2001). For each method (or system) listed, a brief method summary is given. This summary contains a short description as well as evaluation results if available. The methods have been categorised according to the methodologies and resources used in the two steps. For the extraction and the filtering of term candidates internal or external information is used. Internal information can be extracted from the analysed texts themselves and comprises orthographic, morphological, lexical and structural information as well as acronyms and abbreviations (Nenadić et al., 2004b). External information can be acquired from external sources. This can be contextual or statistical information obtained from larger corpora (Spasić et al., 2008), dictionaries, controlled vocabularies, and ontologies.

*Overview on
term recognition
methods
see Table 2.7*

Step 1: Extraction of term candidates

The majority (15/16) surveyed methods analyse natural language text and aim to extract the domain-relevant vocabulary. One method, Lee et al. (2006), composes new terms for the existing Gene Ontology concepts. In the overview in Table 2.7 methods are categorised according to the following characteristics of step 1:

- *Use of linguistic pattern-based approaches:* NPs extraction with patterns for Part-Of-Speech categories
- *Use of linguistic parsers:* NPs extraction using NLP parsers
- *Other approaches:* Term candidates are created from other known terms (NODAL-IDA, FASTR), or terms are extracted as n-grams (Turney, 2003);

Nearly all (14/15) methods generating terms from text rely on linguistic components, which are either Part-Of-Speech tagging or parsing and the selection of noun phrases (NP). In difference to others, the LEXTER system uses Part-Of-Speech categories to identify the boundaries between NPs instead of the phrases themselves. Only Turney (2003) extracts candidate terms as sequences of 1, 2, or 3 words.

In general the extraction of NP in English and many other languages can be performed with high precision and is a state-of-the-art linguistic component. All ATR methods relying in NP extraction perform this in a similar way with varying patterns defined over Part-Of-Speech categories or the different annotated corpora to train the taggers. Available ATR systems like *TerMine* and *TermExtractor* methods reliably retrieve multi-word phrases, but ignore single words. Many biomedical terms are indeed multi-word terms. One indication is that almost 90% of the biomedical

terms in the GENIA³ corpus are compounds (Krauthammer and Nenadic, 2004; Nenadić et al., 2004b). Nonetheless, many important terms are single words – method names are often abbreviated (AFM, PCR, X-ray) and most of biomedical identifiers are single words. These terms are currently not extracted by the reviewed systems. One reason most probably is, that allowing single word terms complicates the filtering (step 2), because a lot of general terminology needs to be recognised and discarded.

Step 2: Filtering of term candidates

With no significant differences in step 1, ATR can be formulated as filtering problem with the goal to rank the domain relevant terms high and remove not relevant term candidates. The methods in Table 2.7 are categorised according to the type of filtering and resources used:

- *Use of linguistic filtering:* Application of linguistic processing like Part-Of-Speech tagging or linguistic parsing, but also specialised approaches using handcrafted syntactic rules to rank or filter terms. The manual creation of syntactic rules works well for small scale examples. A search for terminology on cellular components can be achieved by searching beside others for words ending with “some” aiming to find “endosome” and “lysosome”. A search for cell type names can be achieved by searching for words ending with “blast” aiming to find “osteoblasts” and “cytoblasts”. The drawback of building patterns manually is the limited transferability and scalability.

Nearly all methods (12/16) use or allow linguistic filtering alone or in combination with statistical filtering. The four methods not using linguistic filtering are not primarily extracting terms from text, namely FASTR, Tanabe and Wilbur (2002), Turney (2003), and Lee et al. (2006). FASTR is intended for term normalisation and adds term variants, but does not remove variants. Tanabe and Wilbur (2002) aims to recognise biomedical entities and not general terms. Turney (2003) extracts key phrases not using linguistics, and Lee et al. (2006) also adds novel terms created from other terms but does not filter terms out.

- *Use of statistical filtering:* Frequency counts, co-occurrence, the tf-idf (term frequency-inverse document frequency) weighting, entropy, or statistical tests are used.

11/16 methods use such statistical filtering to rank terms. 8 of the 11 methods rely on simple frequency or co-occurrence scoring. This measures can find frequent, but not the important words in the domain. A contrastive analysis of term frequencies in relation to other domains is a better suitable find the relevant terminology. The ranking of terms is then not only dependent on the text itself, but also from the additional analysed corpora. OntoLearn and TermExtractor do perform such comparisons, Text2Onto at least would allow it.

- *Term normalisation:* Orthographic, morphological, lexical and structural variations of terms are resolved and terms are being appropriately grouped.

³ The GENIA corpus is a manually annotated collection of 2,000 biomedical abstracts (Ohta et al., 2002), in which term occurrences are tagged and further classified using the GENIA ontology. The GENIA resources are freely available at <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>.

2/16 methods explicitly use or evaluate the influence on term normalisation on the performance of ATR methods. Resolving term variation and grouping of terms leads to changes in frequency measures and can lead to improvements in precision of 50% in the extraction of terms and up to 80% for frequently occurring abbreviations (Nenadić et al., 2004a).

- *Disambiguation*: Half the surveyed methods (8/16) disambiguate terms. Two methods use context information in form of words or terms to disambiguate between common English words and biological terms as well as between different technical term in the same way named entity recognition methods distinguish between named entities with same name but e.g. from different organisms. Half the listed approaches also rely on some sort of contextual information. As context, some system extract words surrounding prominent candidates terms and use these words to score all other term candidates. Contextual information can also be extracted from available texts using co-occurrence counts of terms and phrases, both locally and globally, or by directly analysing the syntactic dependencies contained in the domain-relevant text or in existing models in form of databases and ontologies. This disambiguation or interpretation step does involve for LEXTER structural classification in form of learning sub-class relations and in Spasić et al. (2008) the mapping of terms to UMLS semantic types.

Table 2.7 shows that a number of systems have integrated linguistic and statistical filtering with disambiguation achieving either high precision or recall. Few systems deal with term normalisation. The precision in the evaluation of TermExtractor ranges from 0.52 to 0.99. While precision can be estimated with existing vocabularies or user evaluations, the recall achieved by ATR methods is often not given or not clearly described (e.g. NODALIDA). The methods providing recall are designed for document indexing, named entity recognition, term normalisation where it is feasible to objectively judge on correctness and relevance.

Method details for ATR systems and related methods

According to the method overview in Table 2.7 the single methods are in the following describe. For the categorisation each methods most discriminative characteristic was selected. A category was added for ontology learning system which are not restricted to the generation of terms.

Term generation methods specifically using statistical filtering

The listed methods use statistical measures like “frequency-based measures” (e.g., based on absolute and relative co-occurrence frequencies), “information-theoretic measures” (e.g., mutual information, entropy), and “statistical measures” (e.g., chi-square, t-test, log-likelihood, Dice’s coefficient), all methods frequently used in computationally linguistics (Wermter and Hahn, 2004).

TERMS (Justeson and Katz, 1995) The authors of TERMS motivate the difference between terminological and non-terminological noun phrases. The claim that terminological noun phrases (1) have shorter modifiers, (2) are repeated unchanged throughout the document, (3) are enriched in technical text, (4) are composed of

Statistical filtering	Term normalisation	Disambiguation/ Classification	Ontology learning systems
NEUTRAL	FASTR	LEXTER	Text2Onto
TERMS	Nenadić et al.	NODALIDA	OntoLearn
TerMine		CLARIT	
Wermter and Hahn		Tanabe and Wilbur	
		Turney	
		(Lee et al.)	
		TermExtractor	
		Spasić et al.	

Table 2.6. Categorisation of term generation methods

nouns and adjective (if consisting of multiple terms), and (5) the average length is below 2 words. These assumptions have been modeled in a regular expression $((A|N) + |((A|N) * (NP)?)(A|N) * N. .$ Like other pattern based approaches, the method favours precision over recall.

TerMine (Frantzi et al., 2000) / NEUTRAL (Frantzi and Ananiadou, 1995) After NEUTRAL (Frantzi and Ananiadou, 1995), Frantzi and Ananiadou (1997); Frantzi et al. (1998, 2000) developed the C-value/NC-value method. For the ranking of the term this method considers a terms total frequency of occurrence in the corpus, a terms frequency as context word (of the top ranked terms), and the number of top ranked terms the term appears with. Additionally a term is preferred the longer it is by in-cooperation of the length of the candidate string (in number of words) in the measure. Compared to simple frequency measures and depending on the linguistic filter (extraction of phrases based on Part-Of-Speech tagged text) precision increases with the C – value method by 0.06 – 0.08 to 0.40 – 0.44 for those candidate terms which are nested in other terms. For terms which only occur nested precision increased by 0.31 – 0.38 to 0.50 – 0.60). Overall precision increased only 0.01 – 0.02 compared to the frequency measure and reached 0.31 – 0.38. The C-value method is only little depended on the linguistic filter and the method treats term variants as separate terms.

The introduction of the context weighting factor for additional contextual information (NC-value) changes the distribution of precision and leads to an increase in precision by 5% to 0.75 within the top 25% ranked candidate terms. Overall recall is not affected compared to the C-value method as candidate terms are only re-ranked. The C-value/NC-value method is well defined and is suitable to extract meaningful term candidates. One drawback of the method is, that single word terms are being ignored and therefore gene or method names consisting of only one word are not included in the candidate lists. The method is available as web service TerMine, which is available at the National Centre for Text Mining (NaCTeM) located in the UK and is one of the few applications available to the community.

Wermter and Hahn (2006) (incorporation of linguistic and statistical information) The evaluation by Wermter and Hahn motivates the need for the incorporation of linguistic knowledge to outperform pure statistics. The authors show to what extend different methods are capable to re-rank term lists ranked by frequency of

System	Extraction Characteristics			Filtering Characteristics			Description	Quality
	linguistic patterns	linguistic parsers	other	linguistic filtering	statistical filtering	term normalisation		
LEXTER Bourgault (1994)	✓*			✓		✓	*detection of term boundaries; finds head noun in terms; disamb. by structuring/grouping of terms; Engineering (French)	0.95 precision
NEURAL Frantzi and Ananiadou (1995)	✓			✓	✓ freq.		morphosyntactic patterns and list of suffixes; frequency and mutual information, Medicine (English)	0.70 recall
NODALIDA Arppe (1995)		✓		✓		✓	filtering of NPs preceded by determiner or adjective (kind of, some, one, ...; Cosmology (English))	0.95 – 0.98 precision 0.99 – 1.00 recall
TERMS Justeson and Katz (1995)	✓			✓	✓ freq.		term extraction exploits properties of terminological NPs; Chromatography (English)	0.77 – 0.96 precision estim. 0.71 recall
CLARIT Evans and Zhai (1996)		✓		✓	✓ freq.		document indexing; NP extraction, detection of lexical units and grouping; statistical disambiguation; News (English)	0.82 recall
FASTR Jacquin and Liscouet (1996)	✓	✓	✓*			✓	term normalisation: *metarules for morphological variation; reference words provide context; Medicine (French)	0.87 precision 0.75 recall
TerMine Frantzi et al. (2000)	✓			✓	✓ freq.		NP extraction; frequency based filtering; context defining words in the corpus used in ranking; (English)	0.75 precision within top 25% of terms
Tanabe and Wilbur (2002)	✓					✓	Protein name tagging; POS tagger trained to identify gene names using UMLS dict.; context words for disamb.	0.71 precision 0.63 recall
Turney (2003)		✓*			✓ freq.		*extraction of 1-3 grams; filtering by overlap with author asigned keyphrases, co-oc. obtained through internet searches	precision < 0.3 80% sensible terms
Nenadić et al. (2004a)	(terms from TerMine)			✓	✓ freq.	✓	comparison of ATR, term normalisation, manual created rules for variant recognition which lead to improvement of precision by variants/abbrev.	50%/80% improved by variants/abbrev.
OntoLearn: Navigli and Velardi (2004)	✓			✓	✓	✓	system including term, definition extraction, and disambiguation; Tourism domain	0.80 precision 0.55 recall (estimated)
Text2Onto: Cimiano and Völker (2005)	✓*	✓*		✓	✓	✓	framework for ontology learning; provides algorithms for term and relation extraction; * user must select algorithms	37% of users found the system intuitive.
Lee et al. (2006)		✓*					* dependency parsing to find sub-units of GO concepts to create new terms; validation on existing GO after on year	low precision 3.5% added (recall)
Wermter and Hahn (2006)	✓	✓		✓	✓ freq.		comparison two linguistic and one statistical measure used for ATR on their ability to find terminological terms	t-test similar to freq.; improved ling. info.
TermExtractor: Sclano and Velardi (2007)	✓	✓		✓	✓		terms, abbreviations, and formatting (bold, underline, etc.) is considered; use of dictionaries for proper names (gazetteers)	0.52 – 0.99 precision 14 evaluations
Spasić et al. (2008)	(terms from TerMine)			✓	✓ freq.	✓	extraction of terms (Frantzi et al., 2000), disambiguation with UMLS semantic types; metabolomics techniques (English)	two manual eval. 3.5 out of 5

Table 2.7. Overview on term generation systems and their characteristics. All methods use linguistic filtering, most methods statistical filtering, some methods use context information. The quality is given in terms of precision and recall.

occurrence. The authors compare two approaches from the related fields automatic term recognition (ATR) and collocation extraction (CE), the latter one extracts sequences of words co-occurring more often than expected by chance. Limited syntagmatic modifiability (LSM) in CE exploits the linguistic property, that collocations are less modifiable with additional lexical material, meaning that domain specific terms occurring together do not occur together with non-domain specific terms at the same rate. Limited paradigmatic modifiability (LPM) in ATR assumes that domain-specific terms are linguistically more fixed and show less distributional variation, meaning that e.g. the same linguistic component is used in different sentences. They found that statistical based measures e.g. t-tests show no performance difference to the frequency of occurrence counts. They found further, that LPM performs better than LSM, and that all methods promote true negatives instead of keeping them in the lower segments of the ranked list.

Term generation methods specifically using term normalisation

FASTR (Jacquin and Liscouet, 1996) FASTR was developed for term normalisation by detecting term variations. It is capable to extract new term variants from existing validated terms, which at hand restricts the acquisition of new terms where no terms exist. For this purpose FASTR allows the creation of meta-rules to describe common morphological modifications like coordination ("*botulinum type A and B toxins*" → "*botulinum toxin type A and B*"), permutation ("*botulinum neurotoxin type A*" → "*botulinum type A neurotoxin*"), and insertion ("*botulinum toxin A*" → "*botulinum toxin type A*"). Such implicit rules are generalised and meta rules are being dynamically calculated. The meta rules can be re-used.

Nenadić et al. (2004a) (resolving term variation) Natural language texts used as source for ATR methods often contain morphological and syntactic variations of the same term. These variations significantly complicate the process and might influence the results. The correct association of acronyms and abbreviations with the corresponding long forms, the detection of synonyms, or treatment of simple orthographic differences has impact on the frequency values obtained to the candidate terms. Nenadić et al. also discussed the impact of term variations to the ATR results based on an experiment using the TerMine method (Frantzi et al., 2000), additionally considering the following classes of term variation:

- *orthographic*: hyphens, slashes, upper case, lower case, spelling variations and Latin/Greek spelling e.g. "amino acid" vs. "amino-acid", "NF-KB" vs. "NF-kb", "tumour" vs. "tumor", "oestrogen" vs. "estrogen"
- *morphological*: inflection e.g. singular vs. plural; derivation "cell component" vs. "cellular component"
- *lexical*: lexical synonyms e.g. "cancer" vs. "carcinoma" ;
- *structural*: use of prepositions e.g. "clones of human" vs. "human clones"; prepositional variants e.g. "cell in blood" vs. "cell from blood"; term co-ordinations e.g. "human pancreas and liver";
- *acronyms and abbreviations*: "tuberculosis" vs. "TB"

The introduction of inflection variants improved precision by approximately 25%. Acronym variations significantly improved precision by 70% when considering the

most frequent terms and also improved recall up to 25%. Acronym variation detection especially lead to improvement for frequent terms, which are typically abbreviated. The in-cooperation of structural variation negatively influenced precision, because many false positives were introduced. The other variation types had only marginal influence in the ATR results.

Recently Tsuruoka et al. (2008) introduced a framework for the automatic discovery of normalisation rules from dictionaries. The authors evaluated their approach on the UMLS and the gene and protein dictionary BioThesaurus⁴. As one result it was shown, that fully automatically compiled normalisation rules can perform equally well as manually create rules. The results are postulated to improve the performance for term-concept mappings. The rules are extracted iteratively from the gene/protein dictionary. The first five discovered rules were:

1. (*Conversion of capital letters to lower case*)
2. ' ' \Rightarrow ' _ '
3. ' _ ' \Rightarrow ' ' '
4. 'protein' \Rightarrow ''
5. 'precursor' \Rightarrow ''

For the disease dictionary in the first five iterations the following rules have been discovered:

1. (*Conversion of capital letters to lower case*)
2. ' , ' \Rightarrow ''
3. " nos" \Rightarrow ''
4. "[x]" \Rightarrow ''
5. ' o ' \Rightarrow ''

With each iteration the variability of the dictionary decreased. Variability quantifies how variable terms are and is the average on how many unique terms are included by each single concept. The authors are able to show that precision for the look-up of terms only decreases marginally, while recall was improved with each iteration. This means, that grouping lexical variants of terms improves the retrieval performance. The precision stays approximately the same, while recall improves e.g from 0.16 (beginning) to 0.39 (iteration 93) for diseases or from 0.19 (beginning) to 0.41 (iteration 69) for gene/protein names.

Term generation methods specifically using disambiguation or classification

LEXTER (Bourigault, 1994) LEXTER is a term extraction tool and has been developed for document indexing. LEXTER acquires term candidates from a Part-Of-Speech tagged corpus. It uses patterns defined on Part-Of-Speech categories to find boundaries between potential noun phrases. These patterns specify the boundaries between phrases which are pronouns, finite verbs, and conjunctions. LEXTER aims to obtain the phrases of maximal length. The obtained noun phrases are further

⁴ see <http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml>

processed to recursively decompose them into head noun and an expansion part. A structure learning component groups terms together and provides internal context information for the extracted terms. The evaluation of LEXTER as reported by Castellví et al. (2001) lead a very high precision value of 95%. No information on recall has been given.

NODALIDA (Arppe, 1995) NODALIDA-95 is a term extraction which filters obtained noun phrases using morphological and syntactical manually defined rules and disambiguates terms based on only linguistic criteria. The results reported for the noun phrase extraction are very high (precision > 0.95, recall > 0.98) evaluate on a relatively small 20.000 word corpus. It has not been made clear how precision and recall have been obtained.

CLARIT (Evans and Zhai, 1996) CLARIT is a document indexing system, hence has the goal to find indexing terms in text that best describe the according document. The system detects the noun phrases and the lexical atoms within them. In particular, CLARIT deals with the difficulty of grouping nouns phrases appropriately using two heuristics, (1) words that occur together as lexical atom are likely to be used like a single word and (2) if lexical atoms are a noun phrase they hardly allow insertion. With this heuristics noun phrases are sequentially grouped:

OECD validated in vitro test system
OECD validated [in vitro test] system
[OECD validated] [in vitro test] system
[OECD validated] [[in vitro test] system]
[OECD validated] [in vitro test system]

The CLARIT has been evaluated in the TREC-5 document indexing task (Zhai et al., 1997) where it has been shown that replacing the original NLP component by CLARIT lead to similar results. Castellví et al. (2001) reports a recall of 0.82 but also that precision has not been evaluated.

Turney (2003) (Key phrases as source for terms) Libraries and publishers provide access to literature and a major retrieval strategy is based on author assigned keyphases. Keyphrase allow a rough categorisation of a document. As not all documents have keyphrases assigned, the automatic assignment of keyphrases is desired. Turney (2003) compared different features used in keyphrase extraction algorithms.

Lee et al. (2006) (Ontology-centric approaches to ontology learning) While most ATR methods regard text as primary source for new terminology, Lee et al. uses an ontology to predict further ontology terms. Known relationships between concepts are analysed and inherent relationships are inferred to other concepts. The terms “chemokine binding” and “C-C chemokine binding” contain the information on the hypernym relation “chemokine” → “C-C chemokine” which can now potentially become inferred to all concepts containing “chemokine” as a proper substring. By searching scientific literature for sentences or sequences of sentences containing the logical sub-units of a term, the new candidate terms get validated. All terms for which no evidence could be found were rejected. With this approach 3.5% (55/1594) of the

new concepts in GO (Nov.05) could be predicted a year in advance. The idea of the work by Lee et al. is straight forward, but completeness is not necessarily the goal for ontology generation. An overgeneration of terms can be counterproductive, especially in this work where 18,964 candidate terms were generated on the basis of 8,768 terms in GO as of 2004. On the other side especially for complex ontologies with intrinsic naming conventions such as they exist for the Gene Ontology the automatic generation method benefits from this ontology centric approach as it was presented by Lee et al. (2006). Some methods use existing models for prediction of terminologies and for ontology learning in general. Pivk (2006) facilitated web tabular structures and transformed them into knowledge models such as ontologies.

TermExtractor (Sclano and Velardi, 2007) The term extraction described in OntoLearn has been recently made available in another tool called TermExtractor (Sclano and Velardi, 2007), “a web application to learn the shared terminology of emergent web communities.” The tool was evaluated by 14 parties and reached precision results ranging from 0.52 to 0.99.

Spasić et al. (2008) (Re-use of existing controlled vocabularies) Spasić et al. used the method by Frantzi et al. (2000) to extract the terminology relevant in the domain metabolomics. In a small scale manual evaluation 100 terms (out of 1,600 terms) were reviewed of which up to 50% were judge correct.

Disambiguation to finding terms that are named entities The ranking of terms in the biomedical domain can largely profit from the detection of named elements, such as genes and proteins, prior the detection of other terms. This task is called named entity recognition (NER). NER has the aim to locate elements in text which belong to a predefined category. Generally NER includes the extraction elements such as geographical locations, names of personalities and organisations, as well as measures of quantity or time. In biology the most important categories are gene names, protein names or names of organisms. The task goes beyond the simple identification of the boundaries of the element in text. Usually unique identifiers need to be assigned to the found entities. A major problem here is the selection of the correct identifier for a gene or protein. Tuason et al. (2004) reported ambiguities from gene names to general English in the range from 2 to 32% depending on organisms and nomenclatures studied. Hence for named entity recognition of genes and proteins word sense disambiguation plays an important role.

NER methods often use dictionaries for e.g. named entity recognition of gene names (e.g. Tanabe and Wilbur (2002)). Hirschman et al. (2002) gave an overview on the state of the art for named entity recognition methods in 2002 and presented a comparison of 7 methods in Biology. F-measure varied from 0.73 (Collier et al., 2000) for an experiment using Hidden Markov models trained on 80 abstracts and tested on 20 to 0.93 (Fukuda, 1998) in a very small experiment based on 30 abstracts relying on hand-crafted rules. Results on larger data sets are said to cluster between 75 – 80%. An review on the quality of manual curation and the use of text mining for automatic curation of gene products (Winnenburg et al., 2008) also discussed current results on named entity recognition and distinguished between two problems: Automatically recognising a text passage mentioning an entity or concept and identifying the entity itself. They argue that gene name recognition and identification are

difficult, as there is an immense variety of gene names and naming conventions and that human genes have on average 5.55 different names (Wilbur et al., 2007). Gene names literally being the functional or phenotypic description (e.g. p54 meaning the protein has a peak at 54kDa in a mass spectrum) of a gene, as well as abbreviations are especially difficult to disambiguate.

Recently, substantial progress has been made in the field of gene name recognition and identification. The BioCreative challenges (Hirschman et al., 2005; Morgan and Hirschman, 2007) defined benchmark data sets for both tasks in fruit fly, human, mouse, and yeast. The best results for gene name identification range from success rates of around 80% for mouse, human, and fruit fly to over 90% for yeast. For the simpler problem of gene name recognition results are around 87% (Huang et al., 2007; Kuo et al., 2007; Ando, 2007). The best results in the gene normalisation task of the BioCreative II challenge reached an F-measure of 0.86 (Hakenberg et al., 2008). A recent publication from Wermter et al. (2009) reached the same F-measure as Hakenberg et al. (2008) confirming this way the current upper bound.

Term generation with OntoLearn (Disambiguation and the use of WordNet)

Also full ontology learning systems like OntoLearn contain term extraction components. The OntoLearn system extracts a list of syntactically plausible terminological multi-word noun phrases (NPs), like compounds, adjective-NPs, and prepositional-NPs as term candidates. Statistical filtering is applied and the domain relevance is estimated using a combined measure. A term is domain relevant if (a) the term has shows a high entropy in the local corpus compared to other domain specific corpora, and (b) the term is equally used in documents of the target domain (equally distributed). Other than most systems, OntoLearn classifies and disambiguates term candidates. The simple classification uses string inclusion, the concept label is contained in its sub concepts, which can only be found for few candidate terms. For further semantic interpretation OntoLearn uses word sense disambiguation on the basis of on WordNet (Section 2.3.2) and annotated corpora and learns rules for tagging concept pairs with the appropriate semantic relation. By comparing the semantic networks which were extracted from WordNet a meaningful measure to capture the context of a term could be found. The system has been evaluated outside the life science domain to find multi-word terminology in the domain "Tourism" (Velardi et al., 2005). In the evaluation OntoLearn achieved a precision of 0.80 and a recall of 0.55 which was estimated by manually identifying truly relevant terms from a list of syntactically plausible multi-word expressions.

- Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites (Navigli and Velardi, 2004)
- Quantitative and qualitative evaluation of the OntoLearn ontology learning system (Navigli et al., 2004)
- Ontology Learning from Text: Methods, Evaluation and Applications (Velardi et al., 2005)
- Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain (Navigli and Velardi, 2006)

Term generation with Text2Onto (Framework with GUI component) Like OntoLearn, the Text2Onto system (Cimiano and Völker, 2005) is an ontology learning framework which supports the automatic and semi-automatic generation of ontologies from textual documents. It allows the terminology generation from text based on machine learning approaches with basic linguistic processing such as tokenization or lemmatising and shallow parsing. Text2Onto builds on the GATE (Cunningham et al., 2002), a framework and graphical development environment for robust NLP tools and applications. In sequence the following steps are performed to extract terms: tokenization, sentence splitting, Part-Of-Speech tagging, assignment of appropriate syntactic categories, lemmatising or stemming. After the basic linguistic pre-processing pattern-base rules (JAPE rules) can be applied to the annotated corpus to add further annotations. Text2Onto was evaluated by Hatala et al. (2009) who anticipates a number of potential problems. First, the user is not guided in the selection of the available algorithm. One needs to try all combinations and evaluated the results. Text2Onto generates large number of proposed concepts and the users has to individually can accept or reject concepts the suggestions. 37% of the test persons found the process modelled in Text2Onto intuitive.

Terms retrieved from text are grouped in semantic units, often referred to as classes or concepts. The extraction of concepts as reported in (Cimiano, 2006b, p.28) is not clearly defined. Concepts should ideally contain a textual definition and the list of lexical realisations, also referred to as lexica, retrieved from real world texts (Cimiano, 2006b, p.24). Synonyms like “*apoptotic programmed cell death*” for “*apoptosis*” or abbreviations like “*RNAi*” for “*RNA interference*”⁵. For the purpose of text mining syntactic variants with varying case, hyphenation or grammatical number, which are regarded as lexica of a concept should be known. Concepts are identified in text by finding associated terms or simply grouping the terms obtained through automatic term recognition. When grouping concepts the problem of granularity of word senses arises. Different applications need different granularity of senses. E.g. WordNet (Section 2.3.2) is too fine granular for many applications and sense differences are sometimes hard to distinguish. The problem is addressed in “Learning to Merge Word Senses” (Snow et al., 2007).

2.3.2 Finding synonyms

Synonyms are important as authors and annotators may use equivalent, but different terminology for the same concept. Synonyms are an essential source to recognize ontology terms in text, but also to create a mapping between terms existing in different ontologies. For example, authors might refer to the concept fever in different ways. Some texts in Medicine will mention the term “*fever*” itself, others the Latin name “*pyrexia*”. The Gene Ontology synonyms “*apoptosis*” and “*programmed cell death*” are used synonymously in literature. In Go3R, the term “*Bovine Corneal Opacity Test*” has synonym “*BCOP Assay*” with the abbreviated form of “*BCOP*” followed by “*Assay*” as synonym for “*Test*” in this specific context. Other than in the previous examples, terms are often not exact synonyms, but have slightly broader or narrower senses. Sometimes synonyms are even hypernym or hyponyms. For

⁵ examples from the Gene Ontology (March 2008)

Method	Characteristics			Precision/ Recall	Accuracy Confidence	Comment
	patterns	machine learning	resources used			
Landauer and Dumais (1997)		✓			0.64 0.53 – 0.75	evaluated on TOEFL synonyms dataset; using latent semantic analysis
Turney (2001)		✓	AltaVista searches		0.74 0.63 – 0.83	evaluated on TOEFL synonyms dataset; using Point-wise Mutual Information
Jarmasz and Szpakowicz (2003)		✓	Roget's thesaurus WordNet		0.79 0.68 – 0.87	evaluated on TOEFL synonyms dataset
Terra and Clarke (2003)		✓			0.81 0.71 – 0.89	evaluated on TOEFL synonyms dataset; using Point-wise Mutual Information
Turney et al. (2003)		✓	WordNet		0.98 0.91 – 1.00	joint algorithm of previous 4 methods for TOEFL synonyms dataset
Shimizu et al. (2008)		✓		0.19 / NA		using dependency structure
Mccrae and Collier (2008)	✓	✓	WordNet*, UMLS*, Wikipedia*	0.73 / 0.30		binary classification of synsets with six methods, * used for evaluation
Turney (2008)	✓	✓			0.76	TOEFL synonyms dataset; conjointly finding analogies, synonyms, antonyms, associations

Table 2.8. Overview on synonym discovery approaches regarding their characteristics and quality. The best systems achieve high precision (sometimes accuracy was measured). recall is typically lower, meaning that less then half of all synonyms are found.

example the terms “*proof*”, “*finding*”, and “*certification*” are synonyms for the term “*validation*”. The term “*finding*” could be seen as hypernym, “*validation*” is some sort of “*finding*”, and “*certification*” as hyponym, because any “*certification*” is automatically a “*validation*”.

Overview on
synonym
discovery
see Table 2.8

Table 2.8 summarises nine approaches to find synonyms. These approaches can be categorised by their characteristics in using **patterns**, or **machine learning** and by the employed external **resources**. A frequently used benchmark for synonym detection are 80 questions on synonymy of the *Test Of English as a Foreign Language (TOEFL)*. 6 of the 9 reviewed systems compare against this benchmark and achieve a precision between 0.64 and 0.98. All surveyed systems use machine learning to determine synonymy.

WordNet

WordNet is a lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept (Fellbaum, 1998). For example the noun synsets defined for “learning” (Figure 2.7) distinguish between general learning as cognitive process and the type of learning, like “book learning”. The synsets are interlinked by means of conceptual-semantic and lexical relations creating a network of meaningfully related words and

WordNet entry: "learning"	
Noun:	
• S: (n) learning , acquisition (the cognitive process of acquiring skill or knowledge) <i>"the child's acquisition of language"</i>	
• S: (n) eruditeness, erudition, learnedness, learning , scholarship, encyclopedism, encyclopaedism (profound scholarly knowledge)	
Verb	
• S: (v) learn, larn, acquire (gain knowledge or skills) <i>"She learned dancing from her sister"; "I learned Sanskrit"; "Children acquire language at an amazing rate"</i>	
• S: (v) learn, hear, get word, get wind, pick up, find out, get a line, discover, see (get to know or become aware of, usually accidentally) <i>"I learned that she has two grown-up children"; "I see that you have been promoted"</i>	
• S: (v) memorize, memorise, con, learn (commit to memory; learn by heart) <i>"Have you memorized your lines for the play yet?"</i>	
• S: (v) learn, study, read, take (be a student of a certain subject) <i>"She is reading for the bar exam"</i>	
• S: (v) teach, learn, instruct (impart skills or knowledge to) <i>"I taught them French"; "He instructed me in building a boat"</i>	
• S: (v) determine, check, find out, see, ascertain, watch, learn (find out, learn, or determine with certainty, usually by making an inquiry or other effort) <i>"I want to see whether she speaks French"; "See whether it works"; "find out if he speaks Russian"; "Check whether the train leaves on time"</i>	

Fig. 2.7. WordNet entry for the word "learning". The entry shows two senses for learning as a noun and six as verb. Synonyms and a sample sentences is given for each entry.

concepts. Synsets and relations can be accessed and navigated via the provided application programming interface. In 2010, WordNet contains in total 155,287 unique noun, verb, adjective, and adverb strings used in 206,941 word-sense pair. WordNet provides an application programming interface. WordNet captures general English and has no focus on a specific domain. Much of the vocabulary used in biology is not contained. Bodenreider et al. (2003) evaluated WordNet and found that in particular gene product symbols and cellular components are missing.

Machine learning used for synonym discovery

Turney et al. (2003) scores synonymy using web search results and evaluates the approach against TOEFL questions. In the experiment a precision of 0.98 was reached. Drawback of the approach is its computational expensiveness, which make it only suitable for the validation of candidate synonyms and is not a solution for finding synonyms by scanning all possible pairings.

Recently Turney (2008) proposed to unify the algorithms for the recognition of analogies, synonyms, antonyms and associations, which have been treated independently in the past. The supervised learning algorithm is trained for each word pair, e.g. (*term*, *synonym*). The elements of the feature vectors are based on the frequencies of automatically defined patterns in a 250 GB text corpus of web documents. It became evident, that the accuracy of a methodology for finding synonyms corresponds to the size of the corpus used to obtain e.g. pairwise co-occurrences of terms for methods relying on Pointwise Mutual Information or for the learning of patterns of synonymy.

The experiment involved the answering the 80 TOEFL questions. Of 15 previously published results on this task 8 algorithms have higher and 7 lower results as the 76.2% achieved by the joined algorithm compared to 64.5% correct answers for an average applicant to a US university.

Experiment	Accuracy	Best previous	Human	Baseline
TOEFL Synonyms	0.76	0.98	0.65	0.25

Machine learning techniques have been further studied in the context of synonym acquisition from text. Different metrics such as Cosine Similarity, Euclidian Distance, Jaccard Index, Manhattan Distance, Jensen Shannon Divergence, and skew divergence have been compared to machine learning (Shimizu et al., 2008). For the best metric, here a machine learning classifier, a average mean precision of 0.19 was obtained. All other metrics perform even worse.

Patterns used for synonym discovery

Mccrae and Collier (2008) reported for a small scale experiment on learning regular expression patterns for synonymy a low recall of 0.30 at a precision of 0.73 which corresponds to a study in Shimohata and Sumita (2005) reaching 0.21 – 0.27 coverage at a precision greater than 0.7. When checking against WordNet precision was 1 but recall only 0.07. This shows the low coverage of WordNet, which only contained a few requested synsets. The experiment checking the UMLS showed over 0.4 recall at a precision of 0.9.

Hagiwara et al. (2006) investigated the usefulness of word relations, such as sentence co-occurrence, dependency, and proximity and concluded that combinations of several contextual sources lead to more stable results. Further it was experimentally shown, that the results become better the bigger the reference corpus is. The authors determined that word modifications are most significant between all dependency relationships between words. Modification in this context is defined as a limitation or qualification of on word by another word or phrase. In English language nouns or pronouns can be modified by adjectives, while adverbs can modify verbals, adjectives, and other adverbs.

2.3.3 Abbreviation detection

The detection of abbreviations in natural language text is an important task in information retrieval. An abbreviation is the short form of a word or word phrase. Abbreviations are widely used in scientific literature. It can be distinguished between *local abbreviation* and global abbreviations. Local abbreviations are such abbreviations which occur in documents together with their long forms, while global abbreviations do occur without their longforms explicitly stated. As such global abbreviations are often ambiguous, meaning that they correspond to different word senses in different documents. Some examples how local abbreviations occur are listed in Table 2.9. The task abbreviation detection is usually solved in two steps. (1) a dictionary of abbreviation to long form is created, where abbreviations can be assigned to multiple long forms and vice versa. To choose the correct abbreviation

Rule	Example
The first letter of an abbreviation matches the first letter of the meaningful word of the full form.	<i>The Unified Medical Language System (UMLS)</i>
The abbreviation matches the first letter of each word in the full form.	<i>tumour necrosis factor (TNF)</i>
A word in the full form can be skipped if the abbreviation letter matches the first letter of the following word.	<i>extracellular signal-regulated protein kinase 1 (ERK1)</i>
The abbreviation letter matches consecutive letters of a word in the full form.	<i>insulin receptor (InR)</i>
The abbreviation letter matches the last letter of a word in the full form if the letter is an s and if the first letter of the word matches the abbreviation.	<i>cysteine-rich domains (CRDs)</i>
The abbreviation letter matches a middle letter of a word in the full form if the first letter of the word matches the abbreviation.	<i>immunoglobulin G1 (IgG1)</i>

Table 2.9. Pattern-matching rules for mapping an abbreviation to its full form (Yu et al., 2002)

(2) the occurrence of an abbreviation in text has to be disambiguated, meaning that the correct senses (abbreviation / long form pair) need to be selected.

Table 2.10 contains a summary of these methods including performance estimations (if available) and the major characteristics. In literature various methods have been reported to find abbreviations using **machine learning** (Pakhomov, 2001; Nadeau and Turney, 2005; Gaudan et al., 2005; Yu et al., 2007; Okazaki et al., 2008), **heuristic rules and algorithms** (Taghva and Gilbreth, 1999; Wren and Garner, 2002; Yu et al., 2002; Schwartz and Hearst, 2003; Liu et al., 2003; Adar, 2004; Ao and Takagi, 2005; Zhou et al., 2006; Yu et al., 2007; Okazaki et al., 2008) or rely on **statistics** (Hisamitsu and Niwa, 2001; Liu et al., 2003; Zhou et al., 2006).

*Overview on
abbreviation
detection
see Table 2.10*

Recent literature on abbreviation detection methods

Generally, the machine learning approaches which disambiguate “long form”/ “short form” pairs achieve very high accuracy (precision and recall >0.9), however, they require training data. This training data is usually obtained using rule-based methods or available annotated corpora which typically show low recall and high precision. These high quality “long form”/ “short form” pairs are used to train some machine learning classifier using either maximum entropy classifiers (Okazaki et al., 2008), support vector machines (Gaudan et al., 2005; Yu et al., 2007) or Bayesian classifiers (Yu et al., 2007) to find true abbreviations. Considering all approaches in (Table 2.10) abbreviation detection can be regarded as scientifically solved for most domain. Disambiguation using terms from controlled vocabularies (Adar, 2004), context words (Gaudan et al., 2005), or high quality abbreviation data sets improves the results significantly.

The simpler approaches like Adar (2004) extract only acronyms from biomedical literature abstracts. The system achieved a high precision of 0.95 and 0.75 recall on the detection of long form/abbreviation pairs. The long forms were detected by

<i>Method</i>	<i>Characteristics</i>			<i>Precision</i>	<i>Recall</i>	<i>Comment</i>
	rules and algorithms	machine learning	statistics			
Taghva and Gilbreth (1999)	✓			0.98	0.86-0.93	method based on an inexact pattern matching algorithm applied to text surrounding the possible acronym
Pustejovsky et al. (2001)	✓			0.98	0.72	evaluated on Medstract gold standard ⁶
Yu et al. (2002)	✓			0.95	0.70	Rule-based extraction of abbreviations in parenthesis
Chang et al. (2002)	✓			0.80	0.83	evaluated on Medstract gold standard ⁶
Pakhomov (2001)		✓		0.98		acronym detection on 10,000 rheumatology notes
Schwartz and Hearst (2003)	✓			0.96 0.76 0.81	0.82 0.64 0.82	evaluated on Medstract corpus GOLD STANDARD EVALUATION corpus DEVELOPMENT corpus
Liu et al. (2003)	✓		✓	0.9	0.89	extraction of collocations before parenthesis
Adar (2004)	✓			0.95	0.85	evaluated on Medstract corpus
Ao and Takagi (2005)	✓			0.75 0.87	0.63 0.85	evaluated on Medstract corpus EVALUATION corpus DEVELOPMENT corpus
Nadeau and Turney (2005)		✓		0.89	0.88	replication of the algorithm by Schwartz and Hearst (2003) using supervised learning
Gaudan et al. (2005)		✓		0.99	0.98	uses C-Value method by Frantzi et al. (1998) for disambiguation
Chang and Schütze (2006)		✓		0.80	0.83	evaluated on Medstract corpus
Okazaki and Ananiadou (2006)			✓	0.99	0.82 – 0.95	exploits overlapping definitions of acronyms from several authors; evaluated against own corpus
Zhou et al. (2006)	✓		✓	0.97		one third novel and 19% novel non/acronym abbreviations not contained in other databases
Yu et al. (2007)	✓	✓		up to 0.92	up to 0.91	rule-based dictionary construction followed by disambiguation with machine learning
Okazaki et al. (2008)	✓	✓		0.89 – 0.98	0.87 – 0.98	high F-measure of 0.91 – 0.97 depending on the corpus tested; disambiguation is not addressed.

Table 2.10. Overview on abbreviation detection approaches regarding their characteristics and quality. Typically abbreviations can be reliably found using statistics, machine learning or rules (patterns). precision and recall above 0.80 and often above 0.90 have been achieved for various benchmarks.

searching for the longest common sub sequence in conjunction with a set of scoring rules (Taghva and Gilbreth, 1999, see) that favours the first letter of each word of the long form. The algorithms recognises the cases, where the long form precede

the abbreviation in brackets. Morphological similar long forms get merged if the n-grams they contain are similar. Instead of training a machine classifier, common MeSH annotations of the associated abstracts are used to merge long forms sharing the same context.

Extending the approach of Adar (2004), Gaudan et al. (2005) developed a better disambiguation methodology. In contrast to Adar the similarity of long forms is not anymore defined based common MeSH annotations of the abstracts which contain the long forms. MeSH annotations are only available for MEDLINE abstracts and the approach cannot be applied to arbitrary text. The similarity is now defined based on common words contained in the long forms. Acronyms where no long form could be found are disambiguated based on a context model derived from Frantzi et al. (2000). An support vector machine is trained for each sense of an acronyms by incorporating all abstracts containing long forms. Before training the long forms are removed. The authors report to disambiguate acronyms with a precision of 0.99, recall of 0.98, and an accuracy of 0.99.

*Disambiguation
based on
common words*

With a similar approach, the method by Okazaki and Ananiadou (2006) achieved 0.99 precision and 0.82 – 0.95 recall on a self defined evaluation corpus and supports this way the results by Gaudan et al..

The system ADAM, by Zhou et al. (2006) also finds non-acronym abbreviations, a problem which previous systems did not address. Abbreviations are four in a five step procedure with step (1) extracting candidate abbreviations (only single word abbreviations) and surrounding text, (2) identify long forms using statistical information, (3) filter short-form/long-form pairs according to a length ration (≥ 2.5), (4) verifying that short forms are used in text separately from their long forms, and (5) grouping together morphologically similar long forms. ADAM reaches a precision 0.97 and one third of the abbreviations are novel and are not found by other methods, of which 19% of the abbreviations in ADAM are non/acronym abbreviations.

*Non-acronym
abbreviations*

Yu et al. (2007) disambiguate like others, but treats syntactic variations before training. This is said to be especially important when classifying abbreviations in full-text articles. Tested for two machine learning approaches the authors obtained precision/recall for Naïve Bayesian for the Journal of Biological Chemistry (JBC) 0.86/0.79 and for the Journal of Clinical Investigation (JCI) 0.9/0.84. Whereas the support vector machine (SVM) approach reached for the JBC 0.89/0.91 and for the JCI 0.92/0.88.

*Resolving
syntactic
variations
before training*

Motivated by the limitations of manually created heuristic rules to extracted the correct long forms from text Okazaki et al. (2008) proposes an learning approach for the alignment of abbreviations and their long forms.

2.3.4 Generating textual definitions

Creating textual definitions to unambiguously define ontology concepts is one of the time consuming manual processes during ontology development. The automatic extraction of definitions from text is therefore important to ontology learning. Research on definition extraction mainly takes places in the linguistic domain. Especially definitional question answering, which was part of the TREC question answering tasks is highly related. Questions like “What is X?”, “Who is X?”, or “What is X like?”

Klavans and Muresan (2000)	0.87 precision and 0.75 recall for definition extraction
Liu et al. (2003)	0.61 web pages selected which contain definitions.
Westerhout and Monachesi (2008)	$F_2 = 0.71$ for finding <code>is_a</code> patterns in web pages.
Velardi et al. (2008)	0.74 precision with positive on and 0.36 recall (based on 17 terms); 0.85 precision and 96% coverage, terms with at least one good definition (based on 100 medical terms)
Degórski et al. (2008)	$F = 0.30$ for definition extraction

Table 2.11. Overview on the quality of definition extraction and related methods. Early methods show high quality on small benchmark sets. Later methods achieve results below $F \leq 0.4$. Finding `is_a` patterns in web pages perform well. In information retrieval, quality is often measured as F-measure (F), the harmonic mean of precision and recall.

need to be answered by extracting answers to definitional questions from a vast document collection. In contrast to factoid questions or list questions, a definitional question has an expected type of answer, but only the term in the question, which is to be defined. For this specific task, the answers extracted from multiple documents need to be combined in a single answer. This final step is usually very difficult as documents describe the nature of things from different perspectives (Xu et al., 2005).

*Overview on
definitional
question
answering
see Table 2.12*

*Overview on
definition
extraction
methods
see Table 2.11*

Table 2.12 lists six methods for definitional question answering and Table 2.11 lists five general definition extraction methods. The methods can be categorised by the used techniques in

- 7 methods using **lexical** and **syntactical patterns**:
Xu et al. (2003); Yang et al. (2003); Echihabi et al. (2003); Liu et al. (2003); Saggion and Gaizauskas (2004); Han et al. (2006); Storrer and Wellinghoff (2006),
- 1 methods using **grammars**:
Klavans and Muresan (2000)
- 4 methods relying on **machine learning** techniques:
Cui et al. (2004); Westerhout and Monachesi (2008); Velardi et al. (2008); Degórski et al. (2008).

Recent systems like Westerhout and Monachesi (2008) and Velardi et al. (2008) use modern Internet search engines to increase coverage and use machine learning to increase precision. They mine the web with high accuracy patterns find definitions by analysing the text, structure and layout of web documents. The systems achieve a precision above 0.7 with varying recall above 0.3.

For earlier system in particular, the definitional question answering task (Voorhees, 2003) in the TREC 2003 information retrieval competition served as a good evaluation benchmark. The top systems used lexical and syntactic patterns and achieved success rates of over 25%. Their research focuses on extracting definitional phrases which are not full definitions but statements likely to be a necessary part of the terms definition. Usually lexical patterns, such as “*X is_a*” or “*X, such as*” are used to find the definitional phrases. Known difficulties of patterns, such as low recall, will be discussed in Section 2.3.5 (Taxonomy generation). To automatically create or extract definitions the term to define has to be first identified and secondly

Xu et al. (2003)	$F = 0.31$, 1st rank in TREC2003
Yang et al. (2003)	$F = 0.26$, 2nd rank in TREC2003
Echihabi et al. (2003)	$F = 0.27$, 3rd rank in TREC2003
Saggion and Gaizauskas (2004)	$F_5 = 0.24$ for answering definitional questions
Cui et al. (2004)	$F = 0.53$ for answering definitional questions
Han et al. (2006)	$F = 0.16$ for answering definitional questions

Table 2.12. Overview on the quality of definitional question answering. In the TREC2003 task on definitional question answering, the best system achieved a F-measure of $F = 0.31$. Later systems reach up to $F = 0.53$. In information retrieval, quality is often measured as F-measure (F), the harmonic mean of precision and recall.

arising ambiguities have to be resolved. Finally the distinction between definitional and non-definitional phrases has to be performed, because patterns do not occur within definitions only. The sentence (+) below is a true definition and sentence (-) contains the pattern *is_a*, but is not a definition of interest.

- (+) *An **ontology** is a formal representation of a set of concepts within a domain.*
 (-) *We believe that defining the concepts in an **ontology** is a wise investment.*

Since it is easier to reject answers to definitional question than extracting additional missing definitions from the original texts, in definitional question answering recall is usually assumed to be more important than precision in the first place. To judge on a methods potential in an application environment precision and especially rank normalised precision, called average precision (Section 2.2.4) is of importance. The higher the average precision values are the more relevant and correct are the definition candidates presented to the user within an application.

In the following the related work relevant to definition extraction will be presented by summarising the methods, and results for the different approaches and systems described.

Literature on definition extraction using lexical and syntactic patterns

In the evaluation of the TREC2003 question answering task the best performing systems achieved an F-measures F_5 between 0.46 and 0.55. Taking into account that F_5 overweights recall five times over precision the expectation for precision is approximately 0.20. All top systems rely on rules to retrieve definitional statements and use web search results beside the provided corpus.

Xu et al. (2003) the winner in 2003 ($F_5 = 0.55$) retrieved first 1000 documents containing the definiendum, the term to define. Additional kernel facts were extracted from the candidate sentences to be ranked by similarity to the question targets profile using a tf-idf score. Basically the likelihood of facts sharing a context with the question target is used. The question target profile was obtained from known definitions found on web sites. 40 handcrafted rules were used to extract structured patterns that are typically used to define a term. Relation extraction techniques were used to retrieve further relational facts about the question target. $F_5 = 0.55$

Yang et al. (2003) placed second with $F_5 = 0.47$ selected other than Xu et al. (2003) documents containing all words of the term to define. Those sentences form the positive sentence set, all other sentence the negative one. Preceding and succeeding sentences are kept. Anaphoras are replaced by the target and sentences are further ranked using two criteria: (1) sentence frequency; words of an sentence are counted in positive and negative sentence set and a score for each sentence is obtained similar to tf-idf-scores, and (2) snippets retrieved from web search engines using parts of the positive sentences are analysed for the frequency of occurrence of parts of the search target. A *snippet* is a short textual summary typically returned as search result from keyword based search engines. It is extracted from the sections of the whole document which contain the keywords. The sentences are iteratively concatenated till the length limit for the answer is exceeded.

SNIPPET

Echihabi et al. (2003) In the third placed the TextMap system reached an F-measure $F_5 = 0.46$. The system extracts answer candidates from the Web and the given corpus using sentence splitting and a maximum entropy approach to re-rank the candidates. Additionally WordNet glosses, collected biographies and descriptors for proper people as well as a set of subject-verb, object-verb, and subject-copula-object relations are used to score answer candidates. Relations are e.g. relations like "Aaron Copland composed Fanfare for the Common Man", "Aaron Copland was born in 1990".

Liu et al. (2003) also participating in TREC2003 localized definitions in web content. The system was optimized to search for web sites containing definitions using manually collected patterns to find definitional sentences. HTML single structuring elements, such as headings (<h1>, <h2>, ...) or emphasised text, occurring at the top of a page, are assumed to indicate a definition containing document. This strategy seems sensible, as especially lexica, dictionaries, and thesauri show this structure. The hyperlinks on a page were investigated and followed to find further documents containing definitions for the term to be defined. Evaluated on 28 topics the system was able to find on average 61% web pages with definitions within the top 10 results, compared to 0.18 and 0.17 precision for trivial searches with Google⁷ and AskJeeves⁸.

Saggion and Gaizauskas (2004) dealt with the ranking of definition candidates using co-occurring words to better capture the context of a definition and reached a maximal F-measure of $F_5 = 0.24$. As sources for co-occurring words WordNet (Fellbaum, 1998), Britannica⁹ and websites were used and the content was prepared using Natural Language Processing, such as tokenization, sentence splitting to create candidate definition containing phrases.

Han et al. (2006) in difference to previous systems aims to give answers to definitional questions using shorter statements instead of full sentences – which is more difficult. Another difference is that Han et al. tries to judge on the relevance and

⁷ <http://www.google.com>

⁸ former <http://www.AskJeeves.com>, now <http://www.ask.com>

⁹ <http://www.britannica.co.uk>

validity of a candidate statement by obtaining the conditional probability of a statement, under the conditions that each of the external definitions obtained from a dictionary or encyclopedia is also a valid answer to the question. This is an interesting attempt, but the drawback is the generally low F-measure $F_1 = 0.16$, with a recall of 0.34 and a precision of 0.09.

Storrer and Wellinghoff (2006) focused on the selection of definitions by analyzing the defining verb (definitor) to be able to distinguish true definitions from general text. The most common verbs are forms of “to be”, like “is a” or “are”. Especially these forms of “to be” are used equally in definitions and other text. The evaluation lead to 0.31 precision at 0.83 recall based on 80 statements containing forms of “to be”. Overall 0.34 precision at 0.70 recall where obtained for 19 different verbals playing the role of a definitor.

Literature on definition extraction using grammars or parsers

Klavans and Muresan (2000) In contrast to the previously listed pattern-based approaches DEFINDER implemented algorithms using formal grammars for finding definitions. DEFINDER was evaluated in Klavans and Muresan (2001) where the system identified 40 out of 53 definitions obtaining 0.87 precision and 0.75 recall. In a empirical evaluation the author state that especially the usefulness of DEFINDER retrieved definitions and their readability outperforms those definitions found in the UMLS or Online Medical Dictionary. The short paper does not contain enough information to comprehend and judge the method or the evaluation.

Literature on definition extraction using machine learning

The manual creation of patterns for definition extraction is labour intensive and the patterns are often very specific and lack transferability. Machine learning techniques are able to capture the patterns automatically from training examples or allow a classification of sentences as definitions.

Cui et al. (2004) The method uses standard machine learning to select definitions from the TREC2003 corpus but reaches a low precision of only 0.33.

Westerhout and Monachesi (2008) To extract definitions from web pages it is important to how a definition is formulated has to be regarded. Westerhout and Monachesi analysed the performance of machine learning approaches to extract definitions from on web pages formulated in the following five ways:

to be:	<i>Liver is an organ which is present in vertebrates and some other animals.</i>
verb:	<i>The primary spinal tumour affects the spinal cord cell and nerve roots.</i>
punctuation:	<i>Liver: organ present in vertebrates and some other animals</i>
pronoun:	<i>Liver. This is an organ present in vertebrates and some other animals</i>
layout:	<i>Liver</i> <i>Organ present in vertebrates and some other animals</i>

The evaluation of 330 manually annotated definitions allows a quantification of this use of the different types as shown in Table 2.13. Based on this 330 definitions

Type	Number (percentage)
to be	84 (25.5%)
verb	99 (30%)
punctuation	46 (13.9%)
pronoun	46 (13.9%)
layout	7 (2.1%)
other patterns	48 (14.5%)
all tested	330

Table 2.13. Distribution of how definitional statements are formulated from Westerhout and Monachesi (2008). In most 30% of cases the definitional statement is contained in a normal sentence. In a quarter of the cases it is explicitly given with a form of *to be*.

an newly annotated 150 definition they evaluated the performance of the grammar based approach for finding definitions. *to be* types could be found nearly as correct as the learning examples (F_2 0.51 \rightarrow 0.43). Results for the *verb* type on the other side were in the test corpus much lower than in the training data (F_2 0.62 \rightarrow 0.35). The results for the types *punctuation* and *pronoun* were very low. After filtering using a machine learning approach the results for *to be* and *punctuation* type definitions could be increased to $F_2 = 0.71$ and $F_2 = 0.40$.

Velardi et al. (2008) To extract definitions Velardi et al. queries the search engine *Google* for candidate definitions using observed pattern commonly found in definitions. Secondly, Velardi et al. filters these candidates using a machine learning classifier trained based on a training set of > 100 positive and > 50 negative definition sentences from the domains “arts”, “tourism”, “computer networks”, and > 1000 positive and negative sentences from the domain “enterprise interoperability”. The evaluation of 359 predicted definitions from web documents lead to an $F_1 = 0.86$ showing that the system is capable to extract definition from a web document. Velardi et al. evaluates the extraction of definitions for terms previously generated from text with the same system. For 100 medical terms 948 definitions have been generated achieving a precision of 0.85. For 96% of terms at least one good definition was found. It does not become clear under which criteria a definition has been judged as correct. It has also been note by the authors that these results might vary significantly depending on the domain.

Degórski et al. (2008) employs machine learning to classify definitions and uses Part-Of-Speech tags to train the classifier. True improvement only is achieved when “machine learning algorithms are supported by some – relatively trivial – a priory linguistic knowledge”.

2.3.5 Taxonomy generation

For semantic applications the hierarchy of terms or concepts is of great importance. Relationships, such as *is – a* (hypernymy) and *part – of* (meronymy) are defined in a broader sense as subsumption relations of implication which relate to more general concepts in conceptual taxonomies. Subsumption can be defined as follows:

Definition 2.7 (subsumption). *A subsumption defines a lattice (partial ordering) possibly represented as a directed acyclic graph (DAG). In a DAG child nodes may have more than one parent node and hence the graph does not necessarily have to be a tree. The subsumption relation may be seen as a generalisation relation, where the subsumer states a generalisation over the subsumed.*

Table 2.14 lists eight methods capable of obtaining subsumption (taxonomic) relationships. To extract such relationships from text, there have been two classes of approaches described in literature, namely

*Overview on
taxonomy
generation
methods
see Table 2.14*

- **lexico-syntactic methods:** 4 of the 9 listed methods extract relationships from text using manual or learned patterns of hyponymy.
- **statistical methods:** 4 of the 9 listed methods employ statistical measures to determine the existence of taxonomic relationships

Both classes rely on the distributional hypothesis introduced by Harris (1968) which defines that two words which appear in many similar linguistic contexts are semantically similar. Lexico-syntactic methods analyse features of words and how they are composed or modified. Statistical methods analyse the occurrence, co-occurrence, and the distribution of words within and between documents.

In the following section a number of relevant publications are presented. Reflecting the nature of the methods the section is structured in parts for methods relying on **syntactic patterns** and methods relying on **(statistical) similarity measures**.

Taxonomy learning methods using syntactic patterns

Hearst (1992) A first example for lexico-syntactic methods are Hearst-patterns, who compiled a set of lexico-syntactic patterns usually used to describe subsumption (mainly hypernymy) in text. Examples are: *A is a B* or *B such as A*. With these patterns one can infer e.g. from the text fragment “organelles such as mitochondria”, that mitochondria are organelles. To show the wide usage of such patterns the authors analysed corpora and e.g. found in the New York Times news corpus (20 million words) a total of 3178 sentences containing “such as”. Generally it can be said, that the application of Hearst-patterns lead to high precision, but a low recall, since many relationships are not made explicit in text.

Faure and N’edellec (1998, 1999); Faure and Poibeau (2000) proposed a different technique called conceptual clustering. After the acquisition of syntactic frames in a text, the learning method relies on the observation of syntactic regularities in the context of words, for example for such an instantiated syntactic frames is <to travel> <subject: [father, neighbour, friend]> <by: [car, train]>. Concepts found are grouped according to their semantic distance and become this way ordered in a hierarchy. For this, no manual curation is needed beforehand, but the validation of the result is performed manually and is therefore time-consuming. A pattern-based learning approach instead will use labelled examples for extracting instances from texts. While the annotation of the learning examples is time-consuming, the quality of the learning results is be predictable and can be validated automatically.

Method	Characteristics		Comment
	syntactic patterns	statistics	
Hearst (1992)	✓		In an example only 42/3178 (Grolier's Encyclopedia) and 152/7067 (New York Times) sentences which contain "such as" were found to contain a hypernym relation. Hears patterns have high precision (> 0.90) but low recall (< 0.10).
Caraballo (1999)	✓	✓	0.33 use of Hearst patterns; precision (strict), 0.60 precision (by one human judge)
Sanderson and Croft (1999)		✓	co-occurrence measure; no clustering; no learning; 0.48 precision (baseline 0.28)
Faure and Poibeau (2000)	✓		learning of patters for relations from labeled examples
Cimiano et al. (2005)			$F_1 = 0.41$ (Tourism), $F_1 = 0.33$ (Finance)
Snow et al. (2004)	✓		132% improved F-measure compared to classification with WordNet, but generally low maximal F-measure 0.14 (Hearst Patterns), 0.23 (WordNet), 0.27 (TREC hypernyms), 0.33 (TREC hypernyms + coordinate terms), 0.36 (TREC + Wikipedia hypernyms + coordinate terms)
Heymann and Garcia-Molina (2006)		✓	centrality driven creation of noun hierarchies
Snow et al. (2006)	✓		machine learning for patterns using WordNet, TREC, Wikipedia; 0.58 precision, 0.20 recall
Witschel (2005)		✓	co-occurrence in large corpora; 11% to 14% accuracy
Ryu and Choi (2006)		✓*	*review on four methods with recall and precision below 0.50

Table 2.14. Overview on the quality of taxonomy generation. The F-measure is usually below 0.50. In information retrieval, quality is often measured as F-measure (F), the harmonic mean of precision and recall.

Ogren et al. (2004) analysed in an ontology-centric approach to taxonomy generation the compositional structure of Gene Ontology (GO) terms and found that many GO terms contain each other and many GO terms are derived from each other. For example, the term *membrane* [GO:0016020] has *inner membrane* [GO:0019866] as a direct sub-concept. This and similar knowledge can be used to automatically generate new candidate terms following the observed patterns and induce the structure. We evaluated this in a small experiment Section 5.3 (Pattern-based relation extraction – Superstring prediction). Lee et al. (2006) used the taxonomic structure of the ontology to predict new terms including the parent child relations.

Snow et al. (2004, 2006) Instead of creating definitional patterns by hand, machine learning techniques help to learn these syntactic patterns from examples. Snow et al. illustrates that machine learning lead to an improvement of over 132% for finding hypernym relationships compared to simple Hearst patterns. The best setup finds hypernyms with an F-measure of 0.36 and uses training data from WordNet, TREC, and Wikipedia. Secondly, Snow et al. (2006) extend their previous work by an conditional model to judge on the likelihood of generated relations by maximising the conditional probability of relations. A relation is likely to be true if a syntactic pattern exists supporting the assignment. The approach was used to extend WordNet version 2.1 and reports 0.20 recall at 0.58 precision.

Taxonomy learning methods using statistical information

Caraballo (1999) creates hierarchies using syntactic (Hearst, 1992) as well as statistical information, here co-occurrence. The algorithm produces correct hyponyms in 33% of all cases. The evaluation is based on a sample of 10 nodes each dominating at least 20 nouns. The total tree contained 20,014 nouns which have been structured by 654 nodes. Up to three hypernyms were listed as “best” hypernyms for each node. Three human judges had to assess for each noun whether the hypernyms assigned to the corresponding nodes are correct. For 60% of the tested nouns at least one judge judged one hypernym as correct. Given the small test set the evaluation by Caraballo (1999) is not comprehensive. It was not evaluated how many of the nouns within a cluster were correct, just whether the hypernyms assigned to each cluster hold true for the nouns assigned to the cluster. Conclusion drawn by Caraballo are not generalisable for learning taxonomic relations. With the conclusion “.. *that hypernym hierarchies of nouns can be constructed automatically from text with similar performance to semantic lexica built automatically for hand selected hypernyms.*”, the authors compare to the pattern-based approach by Hearst (1992).

Sanderson and Croft (1999) avoided the use of clustering or training data and created concept hierarchies using co-occurrences of concepts (their lexical representations) in text. Half of the pairs obtained by co-occurrence testing fulfill some subsumption criterion.

Heymann and Garcia-Molina (2006) used statistical information for the extraction of subsumption relations from text corpora. In this method two terms are linked if the cosine similarity of their document vectors is above a threshold. The term, which is more central in the whole graph, becomes the parent, the other the child. The *Cosine similarity* is a measure often used to compare text, where the similarity is the cosine of the angle between two n dimensional vectors representing the texts. The algorithm has been described, but not evaluated by the authors. As evaluation for the usage of co-occurrence data this algorithm has been evaluated within this thesis in Section 5.4 (Results: Algorithm by Heymann et. al).

COSINE
SIMILARITY

Witschel (2005) In one of the first large scale evaluations Witschel evaluated to what extent noun phrases can be related via subsumption relations to a hierarchy. The method identifies noun phrases with a pattern based approach using Part-Of-Speech tags, selects candidate terms based on frequency and locates them in a hierarchy by utilising co-occurrence features from large corpora and achieves in the evaluation a low accuracy of 14%. Even though a huge learning corpus (ca. 5 GB) was used the classification data was sparse. Only for 60% of the chosen example, a minimum of 10 similar words could be associated.

Formal concept analysis uses similarity measures to arrange concepts in a hierarchy (see also (Ganter et al., 2005)). On two domain examples *Tourism* and *Finance* the a FCA approach was evaluated by (Cimiano et al., 2005) and compared with KMeans and hierarchical clustering. With $F_1 = 0.41$ (Tourism) and $F_1 = 0.33$ (Finance), FCA outperformed all clustering methods in terms of F-measure. This is due to higher recall values mainly. A drawback of FCA is the exponential time complexity of $O(2^n)$

compared to only $O(n^2)$ or $O(n^2 \log n)$ for KMeans and agglomerative clustering methods. Because partner in a subsumption relationship extracted with FCA can consist of a set of terms, the F-measure is calculated as defined by Maedche and Staab (2002) who uses *Semantic Cotopy*, a measure which averages the similarity between the set of two terms in different ontologies where a set contains all ancestors and descendants of the term. The authors average this over all terms in the learned and the reference ontology. Therefore the F-measure is mostly higher and not directly comparable to methods extracting explicit term-term relationships.

Ryu and Choi (2006) compares four taxonomy learning methods and analysed the features for specificity and similarity in previous methods to select of optimal features to be used for taxonomy learning. Term specificity is a necessary condition for taxonomy learning, because specific terms tend to be located in low level of a domain taxonomy. Term similarity is a necessary condition in taxonomy learning, because similar terms group close together in a taxonomy. Therefore it is highly probable that term t_1 is an ancestor of t_2 in a taxonomy T_D , if both are semantically similar and t_2 is more specific than t_1 in the domain D .

Features for specificity of terms:

- $Spec_{adj}$ – term t (a noun) is specific, if there are few adjectives modifying it (Carballo, 1999; Ryu and Choi, 2005)
- $Spec_{varg}$ – Verb-argument distribution is based on the co-occurrence of terms with special verbs. A term is more specific, if it co-occurs frequently with the same verbs. E.g. "protein" and "increase", "activate", "inhibits", "binds", etc. (Cimiano et al., 2005)
- $Spec_{coldoc}$ – *Conditional probability* of term co-occurrence regards a term t_a to subsume t_b , if $P(t_a|t_b) > P(t_b, t_a)$. Hence t_b is more specific than t_a .
- $Spec_{in}$ – *Inside-word information* is used to measure specificity for multiword terms. Indicates what component word which is highly associated with a term contributes specificity to the term.
- $Spec_{in/adj}$ – harmonised similarity from $Spec_{in}$ and $Spec_{adj}$ to regard both inside and outside information.

Features for similarity of terms:

- If terms co-occur in similar documents, they are similar (Sanderson and Croft, 1999).
- If vectors of adjective patterns of terms are similar, the terms are similar (Yamamoto et al., 2005)
- If vectors of verb-argument dependencies are similar, the terms are similar (Cimiano et al., 2005).

Ryu and Choi compared four taxonomy learning methods and reported recall and precision of 0.50 or lower. It was tested whether the assumption holds, that in a valid parent-child relationship the specificity of the parent is lower than the specificity of the child. While $Spec_{adj}$ showed the highest precision, recall was very low as usually there exist few modifications of nouns by adjectives. Regarding *similarity* it was observed, that taxonomy based similarity ratings are closest to human similarity ratings (correlation coefficient of 0.85).

2.3.6 Availability of ontology generation methods and tools

With the services and web interfaces of TerMine¹⁰ and TermExtractor¹¹ (Sclano and Velardi, 2007), as well as the TerMine Protégé Plug-In¹² there exist methods to extract multi-word phrases from text.

For grouping similar syntactic term variants MetaMap¹³ developed along with the UMLS (Humphreys et al., 1998; Bodenreider, 2004) incorporates a generator of lexical variants, which can be used to join lexical variants of a term for a concept.

From its nature clustering is a intuitive way of grouping terms. Systems like the search engine Carrot2 (Weiss, 2006) or Vivisimo¹⁴, exploit clustering of documents to describe documents by meaningful labels, in fact the labels of the concepts contained in the retrieved documents sets.

WordNet as a semantic lexicon for the English language containing senses for most English words is available for download and can be in-cooperated in applications already.

For biomedical abbreviations, there exist on-line databases like ARGH¹⁵ or the Stanford Biomedical Abbreviation Server¹⁶ which return significance scores for pairs of acronym and long form. The ADAM system¹⁷ (Zhou et al., 2006) reported to have a high precision of 0.97. Other databases like AcroMed¹⁸ and SaRAD¹⁹ are described in literature, but are currently not available.

For definition extraction few tools exist. Search engine provider recently added means to find definitions on the web. www.google.com and Ask.com allow searches for definitions, e.g. `define toxicity`, which retrieves likely definitions for the term toxicity. To do so search engines prioritise rich sources for definitions like dictionaries or lexica. The same resources are used by *GlossExtractor*²⁰ (Velardi et al., 2008) available as web service and helps to extracts glossary entries from text corpora.

Beside relatively precise pattern based approaches, no high quality methods are available.

Ontology Learning Systems

Text2Onto With Text2Onto (Cimiano and Völker, 2005) there exists an ontology learning framework including a graphical user interface which supports the terminology recognition, hypernymic and mereological relationship extraction, and relationship extraction based on statistical significance of linguistically connected text components. It uses a probabilistic ontology model (POM) as representation

¹⁰ <http://www.nactem.ac.uk/software/termine/>

¹¹ <http://lcl2.uniroma1.it/termextractor/>

¹² <http://www.co-ode.org/downloads/protege-x/plugins/>

¹³ <http://mmtx.nlm.nih.gov/>

¹⁴ <http://clustermed.info/>

¹⁵ <http://lethargy.swmed.edu/ARGH/argh.asp>

¹⁶ <http://bionlp.stanford.edu/abbreviation/>

¹⁷ Another Database of Abbreviations in MEDLINE

¹⁸ <http://medstract.med.tufts.edu/acro1.1/index.htm>

¹⁹ <http://www.hpl.hp.com/research/idl/projects/abbrev.html>

²⁰ <http://lcl.uniroma1.it/glossextractor/>

of for the learned knowledge. The POM is a collection of modelling primitives independent from ontology representation languages and are defined in a Modeling Primitives Library (MPL), which contains e.g.:

- concepts (concepts)
- concept inheritance (taxonomic relationships)
- concept instantiation (instances)
- relations/properties (non-taxonomic relationships)
- domain and range restrictions (Axioms)
- mereological relations (part of relations)
- equivalence

Text2Onto provides a number of algorithms to automatically or semi-automatically adapt the POM following the provided the data set.

OntoLT The OntoLT Protégé Plug-in²¹ (Buitelaar et al., 2004) includes rule based extraction of candidate terms and relations based on linguistic features of provided texts. Prerequisite for the extraction process is an annotated corpus of documents as described in Buitelaar et al. (2003). Whether this particular plug-in can be directly utilised in other tasks or not, its attempted to place support directly in a tool used by domain experts is the right way to go to increase user acceptance and finally establish automatic method as part of the ontology creation process.

Ontolearn Several tools emerged from the work on OntoLearn (Navigli and Velardi, 2004). For the extraction of terms there is *termextractor*²², for definitions *glossextractor*²³

Neon Toolkit Recently, the toolkit has been developed as result of the NeOn project funded by the European Union Sixth Framework Programme. The toolkit, as described on the project web site, is “covering a variety of ontology engineering activities, including Annotation and Documentation, Human-Ontology Interaction, Modularization and Customization, Ontology Debugging, Ontology Dynamics, Ontology Evaluation, Ontology Matching, Ontology Specification, Reasoning and Inference, and Reuse.” In particular Ontology Reuse is of interest in this work. The toolkit is based on the Eclipse SDK and builds on Software from Ontoprise GmbH.

²¹ <http://olp.dfki.de/OntoLT/OntoLT.htm>

²² <http://lcl2.di.uniroma1.it/termextractor/>

²³ <http://lcl.uniroma1.it/glossextractor/>

2.4 Summary and Discussion

In the beginning Section 2.1 (Introduction) provided a wider introduction and motivated the importance of ontology learning methods to facilitate the development of biomedical ontologies, to assist biocuration, and to endorse ontology-based literature search. The preliminaries from information retrieval, linguistics, and statistics used in this chapter have been introduced in Section 2.2 (Preliminaries). An overview was provided on the related literature on term recognition (Section 2.3.1) and the associated extraction of synonyms and abbreviations (Sections 2.3.2 and 2.3.3). The literature on the generation of textual definitions (Section 2.3.4) including definitional question answering has been reviewed prior the methods on finding taxonomic relations (Section 2.3.5). The next pages will provide a summary and the critical assessment of the related work from the area of ontology learning which has been presented in this chapter. Specifically, the focus will be set upon the applicability and availability of the published work for the applications in the life sciences motivated earlier. The relation to own work is drawn by identifying open issues required to be addressed for the intended use in ontology engineering, biocuration and semantic search.

Term recognition

In literature the quality of term recognition is describe with precision values between 0.70 and up to 0.98 which truly depends on the experimental setting and the judgement on correctness of predicted terminology. The recall measured was usually above 0.70. The precision depends on how terms are selected from a bigger set of term candidates, often formulated as a ranking problem. High precision corresponds to a good ranking which retrieves domain-relevant terms first. It can be assumed, that for state-of-the-art term recognition a precision of 0.75 and more can be reached within the top region of predicted terms. Hence term recognition is already useful for the fast acquisition of domain vocabulary. For the special case of gene or protein name recognition an F-measure of 0.86 was reached in the BioCreative II challenge. For all evaluations where the terms to extract are contained in the texts, overall recall is only depended on the quality of the detection method for noun phrases. Therefore high recall values can be easily achieved. In open systems without a fixed corpus and therefore no guarantee that requested terms are available, it cannot be assumed that all terms can be found. Here, recall measurements are not representative for a method and are difficult to compare between methods.

*Overview on
term recognition
methods
see Table 2.7*

Concepts, synonyms, abbreviations and lexical variants Concept discovery is not well defined in theory and no tool support is available. Nevertheless it is possible to form concepts by grouping syntactic variations of a term (Section 2.3.6) which then can be enriched with synonyms (Section 2.3.2), abbreviations (Section 2.3.3), and definitions (Section 2.3.4).

In literature synonym detection has been reported to achieve high accuracy with 0.98 for finding TOEFL synonyms (Turney, 2003) and 0.90 precision at a recall of 0.40 evaluated against the UMLS (McCrae and Collier, 2008). Given such results, the methods can be regarded as already good enough to be used. On the other side,

*Overview on
synonym
discovery
see Table 2.8*

the computational expenses of several hours reported by (Shimizu et al., 2008) to use previously learned distances for a word distance based experiment make such approaches not applicable for the interactive acquisition of synonyms. Nonetheless, different senses contained in dictionaries or ontologies can be selected as synonyms after disambiguation.

Overview on
abbreviation
detection
methods
see Table 2.10

Unlike general synonyms, abbreviations (Wren et al., 2005), e.g. the technique “RNAi” standing for “RNA interference” or more precisely for “Ribonucleic acid interference”, can be accurately identified: Gaudan et al. (2005); Yu et al. (2007); Zhou et al. (2006); Okazaki and Ananiadou (2006) report all precision and recall above 0.9. Detection and disambiguation of abbreviations from MEDLINE abstracts can be performed with high quality, and most importantly, there are look-up services available which can be integrated in applications. We conclude that automatic methods can play an important role in finding abbreviations and will most probably be included soon in appropriate ontology engineering tools.

To achieve a meaningful ranking, lexical variations, synonyms, and abbreviations are important. As discussed by Nenadić et al. (2004a), the introduction of inflection variants, acronyms can improved precision significantly. Especially frequent terms typically abbreviated benefit from acronym variation detection. This grouping should also be exploited for the ranking of single-word terms. As shown later, local document frequencies as well as global corpus frequencies significantly change when grouping different variants of terms, compare Section 3.1 (DOG4DAG Term Generation Method). With respect to synonymy, abbreviations and lexical variants the result of the literature study is, that the robust discovery of synonymy is very subjective as synonymy can be defined in various ways depending on the intended use. Promising methods exist, require big amounts of data, and are usually computationally very expensive. Therefore synonym extraction is not applicable for the intended use scenarios. The task of finding abbreviations and lexical variants can be regarded as scientifically solved, but few available implementations exist. For simple abbreviation detection the method of Adar (2004) is sufficient. In cases where disambiguation is required the method of Gaudan et al. (2005) could be used.

For the extraction of term candidates it becomes clear from the literature and explicitly formulated by Wermter and Hahn (2006), that linguistic information matters. Part-Of-Speech tagging or parsing should be used for extracting candidate phrase, e.g. noun phrases. Wermter and Hahn also highlighted in his analysis, that all methods, no matter if they are based on, statistical test, ATR, or collocation extraction, promote true negative terms instead of keeping them in the lower segments of the ranked list. This suggest that improved methods should build on better comparative measures using stable corpus statistics to normalize local observations.

As done for evaluation by Lee et al. (2006), existing ontology terms should play an important role in the validation of term candidates extracted from text. Extracting terms with definitional character from text as they exist in the GO (Ogren et al., 2004) (see also illustration in Figure 2.4) has not been addressed in literature.

Applicability of term generation tools The literature overview describes methods and available tools to extract terms from text. For example TerMine (Frantzi et al., 2000), or TermExtractor (Sclano and Velardi, 2007) are the two available tool which

are well defined and lead to the extraction of meaningful multi-word term candidates. A drawback of the methods is, that single word terms are being ignored and therefore gene and method names, topic descriptors and substances identifiers consisting of only one word are not included in the candidate lists. All methods aiming to extract domain-relevant vocabulary extract terms from text composed of at least two words. This is reasonable, because it is difficult to distinguish between domain-relevant and non-domain-relevant single word terms because corpora usually contain less domain-relevant terms than not domain-relevant terms. Also, multi-word terms in English are a priori most likely to be a technical term and hence can be easily extracted.

Based on result from in literature it can be expected that more than 75% of the terms presented to the user by a automatic system are domain relevant terms. It remains open work how to extract and rank single and multi-word terms by relevance to the domain of interest. In this thesis a method for term generation has been specified, implemented and evaluated (Chapter 3). This method recognises local abbreviations and lexical variants found in text. A ranking method has been specified and evaluated which allows to extracts domain relevant single and multi-word phrases in on result set.

Definition extraction

Non of the reviewed systems is capable to generate complete well structured definitions fully automatically. Definition extraction as required for ontology development should retrieve a selection of proper definitions for terms. Results by Liu et al. (2003) and Westerhout and Monachesi (2008) show that textual definitions can be found on web pages with $F_1 = 0.71$. But, different to Liu et al. (2003), the task now is not ranking web pages, but ranking definitions. The extraction of existing definitions from text can be performed with a reasonable precision (above 0.7) but low recall (below 0.4). The user defined benchmarks used to measure performance are not comparable between systems as they vary strongly in size and the expected quality.

Overview on definition extraction methods see Table 2.11

Extracting definitional statements has been done as part of definitional question answering. Definitional question answering can be an indicator of the state-of-the-art as the task is to retrieve sentences from a corpus which contain the definitional statement. This does not mean, that necessarily a proper definition of a term need to be obtained. Definitional question answering has been evaluated for domains other than the life science. The evaluation of definitional question answering systems in TREC 2003 lead to comparable results. The results for definitional question answering in the literature reach F-measure values between $F_1 = 0.16$ and $F_1 = 0.53$. The best systems in TREC2003 achieved F-measures around 0.30, later systems have been reported to reach $F = 0.53$ (Cui et al., 2004). The extraction of shorter statement lead to lower performance like Han et al. (2006) with $F = 0.16$. The results suggest that it is advantageous to use (1) context information for disambiguation, e.g. co-occurring words, (2) external information, e.g. WordNet, web pages, for re-ranking and validation of candidate statements, and (3) handcrafted rules to perform in TREC definitional question answering. Not unimportant for good results is the size of the answer in the question answering task. Good systems should be able to estimate correctly, which definitions are likely to be correct and of relevance.

Overview on definitional question answering see Table 2.12

Recently, Velardi et al. (2008) evaluated for few terms the extraction of definitions from web pages and presented a F-measure of 0.86. This score was calculated for the whole set of generated definitions (also multiple per term), regardless whether for each term a definition could be found. Hence, multiple, easy to find, good definitions for widely used domain specific terms will influence the performance measurement positively, while terms where it is difficult to find definitions have a lower influence on the result. Finding definitions is especially difficult for new domain specific terms – which are of special interest when creating ontologies – and for general terms which are likely ambiguous and often used. Recall should be calculated weighing all terms equally.

From literature it remains open how well domain specific textual definitions can be extracted in a life science domain. The goal is to find proper definitions of the form A is B with property C. To establish a method, the quality of the method has been evaluated. In this thesis a method for definition extraction from web search results has been developed (Chapter 4). The method has been manually evaluated on the common benchmark from question answering from TREC2003. More importantly the method has been evaluated large scale in the life sciences by manually validating generated definitions for 1,000 randomly chosen terms from MeSH and GO.

Taxonomic generation

*Overview on
taxonomy
generation
methods
see Table 2.14*

There are lexico-syntactic and statistical methods to extract taxonomic relationships such as is-a and part-of from text. In general, the most challenging problem for such methods is the fact that many relationships are not made explicit in text. An example for a lexico-syntactic method are Hearst-patterns (Hearst, 1992), like X such as Y. With these patterns one can infer e.g. from the text fragments “organelles such as mitochondria”, that mitochondria are organelles. Pattern-based methods show typically high precision around 0.90, but a low recall of 0.10. An example for the statistical methods is reported in Heymann and Garcia-Molina (2006). Here, the decision on the existence of a relation between two concepts depends on the measured co-occurrence of the concepts. The concept that shows more independence of the other will be suggested as parent in that relation. Another approach, which can generate ontologies from the concept usage in documents, is formal concept analysis (Ganter et al., 2005; Cimiano et al., 2005) reaching an F-measure of 0.39 – 0.45. Generally statistical approaches reach a F-measure of below 0.50 (Ryu and Choi, 2006; Snow et al., 2006). Recent work by (Brewster et al., 2009) describes an experiment of creating an initial hierarchy of terms in the animal behaviour domain. Like Hearst they use lexico-syntactic patterns to obtain taxonomic relations. The authors provide an small evaluation for 198 selected out of 13.755 terms between which predict relations with an precision of 0.28. This suggest, that simple string inclusion as used is not suitable to obtain good taxonomic relations.

To judge on the applicability of automatic methods for the determination of hierarchical information it has to be clarified how often hypernym relationships can be found in text. Snow et al. (2004) gives evidence, that hypernym and non-hyponym word pairs were found by a ratio of approximately 1:50. Generalised methods can rely on machine learning as done by Snow et al. (2004) who could reproduce that all patterns originally suggested by Hearst

(1992) where also identified by the proposed machine learning method to find patterns for hypernymy. Nonetheless, with recall far below 0.50, taxonomy generation remains an open problem.

Ontology learning integration in ontology editors

Apart from the TERMINE term recognition plug-in for Protégé, non of the currently used ontology editors comprises automated support for finding term, definitions, or novel taxonomic relationships.

To address the needs of the ontology engineers who develop the steadily growing number of ontologies in Biology and Medicine, the ontology generation methods developed in this thesis have been integrated in the two most used editors OBO-Edit and Protégé (Chapter 7)

Terminology Generation

References

Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. In *ISMB (Supplement to Bioinformatics)*, Impact factor 2009: 4.3 (accepted for publication)

Alexopoulou, D.* , Wächter, T.* , Pickersgill, L., Eyre, C., and Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*, 9(Suppl 9):S2, Impact factor 2009: 3.7 *shared first author

Wächter T., Tan, H., Wobst, A., Lambrix, P., and Schroeder, M. (2006). A corpus-driven approach for the design, evolution, and alignment of ontologies. In *Proceedings of Winter Simulation Conference (Computational Systems Biology)*, Monterey, CA, USA (3rd–6th December 2006), pages 1595–1602, Invited contribution.

A term generation method has been designed, implemented, and evaluated. It generates terms by identifying statistically relevant noun-phrases in text and leads to high quality terms also found in manually created ontologies. It specifically deals with the relevance ranking of single word terms by the use of term normalisation and reference corpora which are big enough to capture the relative frequency of terms in comparison to each other. The method was evaluated manually by domain experts and against an ontology which has been created in collaboration with Unilever Research UK in the domain of lipoprotein metabolism. The method outperforms other systems and is capable to retrieve 75% relevant terms in the top 50 and 55% in the top 200 generated terms. Further, it has been analysed how the ranking of terms depends on the selected Part-Of-Speech tagger for noun phrase extraction, the source for the global corpus statistics as well as on the used scoring method. Especially the use of a reference corpus has a high influence on the ranking of terms, but surprisingly the differences are not significant whether a general corpus based on web sites or a domain specific corpus like PubMed is used to obtain global term frequencies. The method has been encapsulated in a web service which enabled the integration in various applications.

From our experience the manual collection of domain relevant terms is time-consuming and not objective. In need of faster acquisition of domain relevant terminology, automatic term recognition methods have been already developed in the 1990s (Section 2.3.1) and are now again of interest to efficiently create vocabularies and ontologies for semantic applications. With respect to **Research Question 1** – *To what extent can ontology construction be automated?* (Section 9.1) – a term recognition method has been developed within this work to automatically generate terminology from natural language texts and to rank these terms by relevance for easier affirmation by a domain expert. In contrast to other existing methods, the new method explicitly allows the extraction of single word terms, which significantly complicates the relevance ranking. The majority of technical terms i.e. almost 90% of the biomedical terms in the GENIA text corpus are compounds. The majority of compounds in a corpus will automatically be domain relevant. In English, there are much more single word nouns than compounds and only a small fraction will be domain relevant technical terms in the domain.

In biology and medicine, single word terms are required to be extracted because e.g. many identifiers, names, acronyms, cellular components, species, or anatomical terms are single words which need to be contained. To clarify to what extend the ranking of terms changes with different configurations of the term generation pipeline and to check the hypotheses associated with research question 1, a number of experiments have been performed. In particular it has been experimentally investigated how the impact of technical parameters such as the specific method to assign part of speech tags (Section 3.5.1), the source of corpus statistics (Section 3.5.2), and the chosen statistical measures (Section 3.5.3) influence the obtained ranking of candidate terms.

Requirements

With respect to the intended use in biomedical applications a number of requirements have been collected:

- (1) The method should represent the state of the art, and as such perform comparable well or better compared to other methods.
- (2) The method should favour terminology from the biomedical domain.
- (3) The method needs to be fast enough to be used in interactive applications, e.g. ontology editors or on the fly document classification.
- (4) The method should be organized modular, to allow to experiment with different algorithms and components.

Requirement (1) Despite the availability of other term generation (or term recognition) methods (Section 2.3.1) a different method was needed to support the process of developing novel biomedical ontologies e.g. to be used in semantic text-mining based applications like GoPubMed. From the few existing systems, TerMine (Frantzi et al., 2000) for instance is available as ready-to-use web service, but ignores all phrases consisting of one word only. The same applies to TermExtractor (Sclano and Velardi, 2007). Both systems build on the assumption that longer terms have on average a higher probability to be technical terms. This assumption certainly holds and while neglecting single words both systems achieve good performance. But single

word terms are very important to be extracted as they often have a distinct meaning in the domain. Many gene, protein, species names, general anatomical terms, and method names are single words and must not be ignored.

To allow appropriate ranking even after including single words the normalisation of term variation is important. As shown in Nenadić et al. (2004a), the introduction of inflection variants improved precision by approximately 25%. Acronym variations significantly improved precision by 70% when considering the most frequent terms while recall improved up to 25%. Acronym variation detection lead to improvement for frequent terms, which are typically abbreviated. The new term recognition method in-cooperates such information on abbreviations and acronyms, as well on lexical variants found in the analysed text. Additionally local and global corpus statistics were in-cooperated to be able to rank single word terms appropriately. The occurrence and co-occurrence counts used to score the candidate concepts (tf-idf or probability based) are being accumulated for all lexical variants associated with a candidate concept. As Liu et al. (2002) found that 80% of the abbreviations defined in the UMLS have ambiguous occurrences in MEDLINE, this method resolves local abbreviations only. If an abbreviation is known, but not contained in the local set of documents the term and its abbreviation are regarded as separate terms.

Requirement (2) To archive awareness and good term rankings for biomedical domain, occurrence and co-occurrence data was extracted from a large corpus which are in the case of PubMed approximately 19M scientific abstracts. This makes the method domain dependent.

Requirement (3) To meet the runtime requirement, an efficient data structure to for text, the TextTree (Section 6.1), has been developed. It allows index based nested annotation of text in linear time.

Requirement (4) To achieve modularity the concept of taggers and revisions has been introduced. The taggers (Section 6.2) assign attributes to text ranges within a text. Attributes can be any type of information. Taggers are technically independent from each other. Hence the text-mining algorithms are specified in separate modules. Nonetheless taggers can depend on each other, e.g. a tagger for noun phrases will depend on a previously applied tagger performing sentence splitting. The revisions (Section 6.3) modify sets of concept e.g. merge, filter, or enrich them with information.

3.1 DOG4DAG Term Generation Method

The *Term Generation Method* was developed to meet requirements described above. Together with definition extraction and taxonomy induction the collection of tools and applications will in the following be referred to as *Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG)*. A basic assumption underlying the term generation pipeline is the one on the morphology and importance of terminology. It has been assumed that terms are statistically relevant noun phrases, which are the units of text possibly being subject or object in a sentence. The judgement on the importance of such phrases is made according to the occurrence statistics of a phrase in the selected document or document set (the one from which the terms are extracted) in relation to the occurrence statistics of the phrase in a domain specific or general reference corpus.

Method summary

We extract terms from English text, which we tokenize before POS-tagging, sentence identification, noun phrase and local abbreviation detection. As POS tagger we use Ling-Pipe-Tagger¹ trained on MEDLINE and the TNT tagger (Brants, 2000) trained on the Wall Street Journal corpus. We generally regard phrases with pattern $[adj|verb] * [fill]\{2\}[noun]^+$ as noun phrases, where *fill* are fill words like *of*, *the*, *for*, etc.

We first great for each noun phrase a candidate concept. In *DOG4DAG*, all lexically overlapping concepts are also grouped. Two concepts overlap, if any two of their lexical representations are similar. They are regarded as similar if they show a Hamming distance of less than 20% of the length of the shorter term label. The Hamming distance between two strings denotes the number of position the two strings differ from each other. We align strings from the beginning and include the length of overlapping tails in the distance. This grouping is not performed if concepts only have common abbreviations. Nested terms, i.e. noun phrases within longer noun phrases, are expanded as separate candidate concepts. *DOG4DAG* not only retrieves the maximal length terms, but also nested terms that follow noun phrase pattern specified above.

We rank candidate concepts, under consideration of all lexical variants and abbreviations, according to their relative importance by the tf-idf (term frequency-inverse document frequency) measure, a weighting method commonly used in information retrieval. It captures the importance of a term in a set of documents in relation to a corpus. As corpus we used all scientific abstracts listed in PubMed.

The method with all modules (taggers and revisions) is illustrated in Figure 3.1. As representation for text a tree structure is used where each node represents a non-overlapping sub-string of the text. Nodes correspond to tagged regions, e.g. tokens, sentences. Nodes can hold several types of tags, e.g. tokens will have Part-Of-Speech tags if they represent a word (Section 6.1).

In the following the steps of the term recognition pipeline (Figure 3.1) will be shortly described in the same order they are applied:

¹ alias-i.com/lingpipe

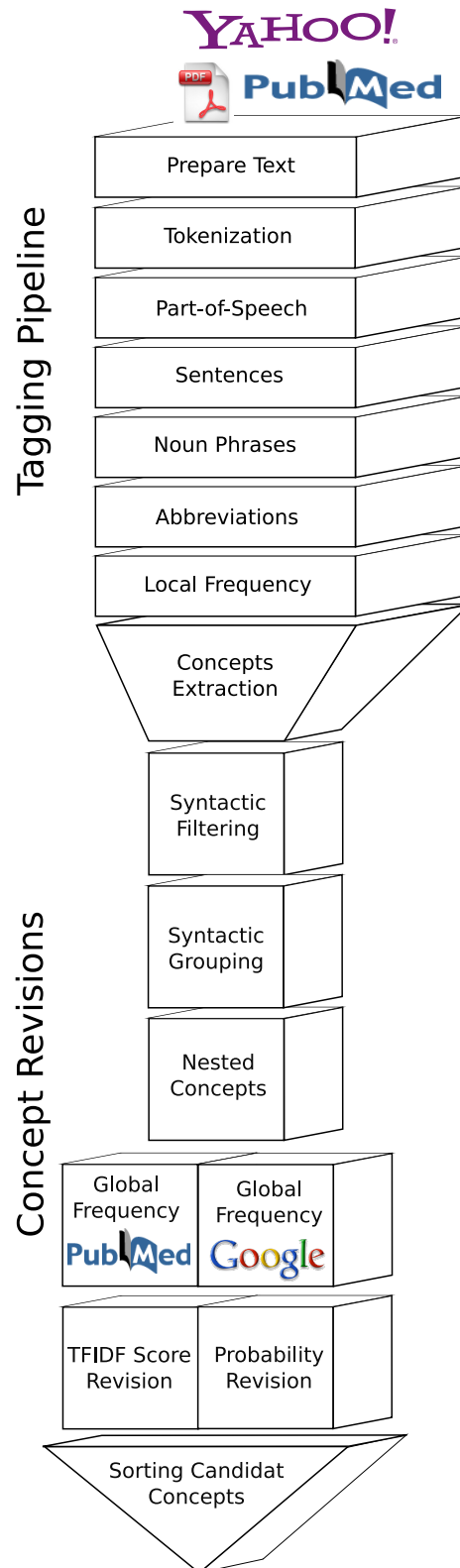


Fig. 3.1. The term generation pipeline

The term generation pipeline

1. **Prepare text:** The initial step in the term recognition pipeline is the extraction of text. As input for the term generation the text from PDF documents, the snippets from Yahoo search results, and title and abstract from PubMed documents is used. The texts are converted in the TextTree representation (Section 6.1), which is the data structure that can be efficiently manipulated by the various ElivagarTagger (Section 6.2) implementations.
2. **Tokenization:** The tokenizer separates text in lexical units, such as word tokens, white spaces, (opening/closing) brackets, punctuations, (opening/closing) quotes. Each of those will be marked up.
3. **Part-of-Speech tagging:** Part-Of-Speech denotes the grammatical classification or the category a word can be assigned to in the context of a phrase, sentence or paragraph. This categories contain classes like noun, adjective, adverb, verbal. For each word token the Part-Of-Speech category gets identified and marked up.
4. **Sentences segmentation:** Sentence borders are identified to structured text as series of sentences.
5. **Noun phrase tagging:** Depending on the used Part-Of-Speech tagger noun phrases are being identified and marked-up following defined consecutive patterns of Part-Of-Speech tags.
6. **Abbreviation tagging:** Local and common abbreviations are being identified and marked up. To find local abbreviations we adapted an implementation from Schwartz and Hearst (2003) to find the best long form.
7. **Local frequency:** The frequency of occurrence is obtained for each noun, noun phrase and abbreviation.
8. **Concept extraction:** Concept representations are created for the nouns and noun phrases obtained from text.
9. **Syntactic filtering:** Concepts are filtered to exclude known false predictions from the result set. Those include document type specific vocabulary, as “Introduction”, “Methods”, “Conclusion”, etc., but also wrong classification of words as nouns by the Part-Of-Speech tagger.
10. **Syntactic grouping:** Concepts are grouped together if they share lexical representations.
11. **Nested concepts:** Concepts get derived from composed noun phrases. Each single noun phrase is added as candidate term.
12. **Global frequency:** From big corpora, i.e. all approximately 18M PubMed abstracts, the frequency of occurrence is obtained for each noun, noun phrase and abbreviation associated with a concept.
13. **Probability revisions (PVALUE):** Given the local and global frequency values the probability of a concept occurring in the local document set is calculate and assigned to a concept. A hypergeometric distribution of random variable was assumed (for details see Section 3.5.3 (Hypergeometric distribution)).

14. **TFIDF revision:** Given the local and global frequency values the tf-idf value is calculated and assigned to a concept.
15. **Sorting candidate concepts:** Candidate concepts are ranked by the scores obtained from the probability (*PVALUE*) or *TFIDF* revisions.

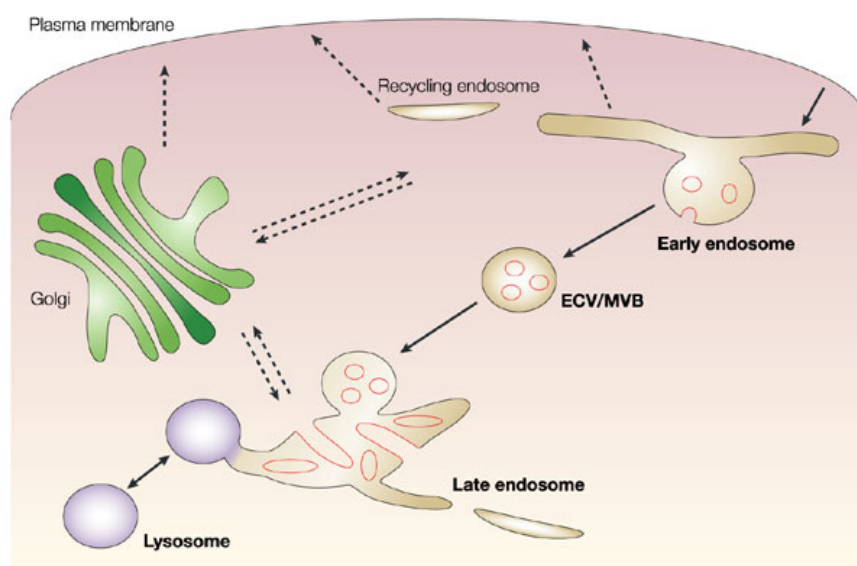
Technical implementation

The whole term recognition pipeline was implemented using the programming language JAVA and was encapsulated in an Axis2 generated web service allowing synchronous and asynchronous requests. This allows the seamless integration into other application as done for the web-based Term Generation Platform (Section 7.5), the OBO-Edit Ontology Ontology Generation Plug-in (Section 7.2), the Protégé Term Generation Plug-in (Section 7.3), and the Go3R Ontology Editor (Section 7.4). The generation of terms based on 250 PubMed abstracts takes approximately 1-2 seconds.

As source for PubMed abstracts the GoPubMed internal search API was used. The web service is running on an standard application server with 4 cores and 4GB of main memory together with the services for Definition Generation and Ontology Look-up. The algorithms are modularised and share the not definition specific components, like tokenization, noun phrase extraction, DOM like data structures, abbreviation detection etc. with the Definition Generation module. All system components are managed using MAVEN 2.0, a software project management and comprehension tool. The module was developed using the Open Source IDE Eclipse.

Applications

The method has been integrated in several applications, among other the two most used ontology editors OBO-Edit (Section 7.2) and Protégé (Section 7.3). The other tools are the Go3R Ontology Editor (Section 7.4) and the web-based term generation platform (Section 7.5).



Nature Reviews | Molecular Cell Biology

Fig. 3.2. Illustration of the endosome biogenesis (from Jean Gruenberg & Harald Stenmark. *The biogenesis of multivesicular endosomes*. *Nature Reviews Molecular Cell Biology* 5, 317-323, April 2004) Endocytosis is the process by which cells incorporate materials from the outside of the cell. The early endosomes located at the cell membrane receives vesicles coming in the cell and releases endosomal carrier vesicles (ECVs) or multivesicular bodies (MVB) which are further received by the late endosome. The late endosomes pass the material on to lysosomes.

3.2 Proof of concept

An initial experiments show that the term generation results are corresponding to the human perception of the importance of terms in documents set. We show on the example of the research topics of senior researchers and group leaders in the BIOTEC and MPI-CBG that term generation retrieves relevant terminology describing the research topics well. To illustrate the nature of the task to be solved by ATR see the following examples 3.1 and 3.2:

Example 3.1 (Term recognition and relevance ranking in molecular cell biology). As example we generated the terminology for the topic endocytosis, the field of work of Mario Zerial, currently director of the MPI-CBG in Dresden and compare against the work description on Marino Zerial's web site (as of September 2008) that is:

*".. molecular mechanisms of **endocytosis**, which is an essential function of all eukaryotic cells. He is specifically interested in the mechanisms underlying **endosome** biogenesis, wants to explore how endocytic **transport** regulates and is modulated by intracellular signalling and in the regulation of endocytosis in polarised cells, such as epithelial cells and neurons."*

The extracted terminology contains chemical compounds, cellular components, laboratory techniques, species, protein or gene names, etc. The ranked list of extracted candidate terms shown in Table 3.1 illustrates how ATR methods select the

1	Rab5	<i>gene/protein</i>
2	endosomes	<i>cellular component</i>
3	Rab	<i>protein family</i>
4	early endosomes	<i>cellular component</i>
5	effector	<i>general term</i>
6	GTPase	<i>chemical compound</i>
7	phosphoinositides	<i>chemical compounds</i>
8	Rab4	<i>gene/protein</i>
9	membrane	<i>cellular component</i>
10	endocytosis	<i>biological process</i>
11	Htt-associated protein 40	<i>protein</i>
12	Rabenosyn-5	<i>protein</i>
13	Huntingtin	<i>gene</i>
14	transport	<i>process (biological?)</i>
15	motility	<i>physical property</i>
16	recycling	<i>process (biological?)</i>
17	domains	<i>protein domain</i>
18	Rabankyrin-5	<i>protein</i>
19	small GTPase Rab5	<i>protein</i>
20	Huntington's disease	<i>disease</i>

Table 3.1. Term list for “endocytosis”. Terminology extracted from all abstracts listed in PubMed for the query *Zerial[au] Dresden[ad]* for a research group working on endocytosis. Overlapping terms with Marino Zerial's research description are shown in boldface.

most important terminology from text. With the terms “*endosome*” or “*endosome biogenesis*”, “*endocytosis*”, and “*transport*” the work of Marino Zerial is described on his groups web site. ATR retrieves these terms and the genes and proteins the group works on in position 2 (“*endosomes*”), 10 (“*endocytosis*”), and 14 (“*transport*”). Additionally terminology related to “*endosome biogenesis*” (see Figure 3.2) is retrieved at position 4 (“*early endosome*”), 9 (“*membrane*”), and 16 (“*recycling*”).

Example 3.2 (Research topics of BIOTEC and MPI-CBG groups). To extend this, the experiment has been repeated for 7 other group leaders. PubMed abstracts are retrieved by a query for the corresponding group leaders name followed by an expression to filter by affiliation, namely *Max Planck[ad] AND Dresden[ad]*. In the following there are the top 10 ranked candidate concepts displayed for each of the 7 group leaders from the MPI-CBG (Example 3.2) (see tables 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8). It was tested whether these terms occur in the short description of their work as given on the groups web pages (retrieved: 30th July 2009). Terms which occur on the web page are displayed in **bold** face. Terms which occur on the web page but occur as part of a longer term only or are not significant for the description of the scientists work are displayed **bold and italic** face.

The vast majority, namely 68/80(85%) of terms extracted from the publications of a author appeared on the web pages. 57/80(71%) have been judged to be significant to describe the authors research. This simple test shows that the extraction of vocabulary from PubMed abstracts is feasible and the top ranked vocabulary is relevant within the biomedical domain.

Label	Abbrev.	Lexical variants
centriole	MT MTs	centriole Centrioles centrioles
microtubule		Microtubules microtubules microtubule
centrosome		centrosomes centrosome Centrosomes
XMAP215		XMAP215
<i>elegans</i>		elegans
<i>assembly</i>		assembly
<i>polarity</i>		polarity
embryo		embryos embryo
spindle	RNAi	spindle spindles
RNA-mediated interference		RNA-mediated interference

Table 3.2. Top ranked terms for Hyman Group

Label	Abbrev.	Lexical variants
morphogenesis		morphogenesis
<i>cell</i>		cells cell
movement		movement movements
embryo		embryos embryo
vertebrate gastrulation		vertebrate gastrulation
gastrulation		Gastrulation gastrulation
slb/wnt11		slb/wnt11 Slb/Wnt11
zebrafish		zebrafish Zebrafish
Wnt11		Wnt11
zebrafish gastrulation		Zebrafish Gastrulation zebrafish gastrulation

Table 3.3. Top ranked terms for Heisenberg Group

Label	Abbrev.	Lexical variants
esiRNA		endoribonuclease-prepared siRNAs esiRNA esiRNAs endoribonuclease-prepared short interfering RNAs
screen		screen screens
gene		gene genes
C13orf3		C13orf3
cell division		Cell division cell division
RNA		RNA RNAs
<i>cell</i>		cells cell
RNA interference		RNA interference
AML1		AML1
short interfering RNA		short interfering RNA short interfering RNAs siRNAs

Table 3.4. Top ranked terms for Buchholz Group

Label	Abbrev.	Lexical variants
bounds		bounds
microtubule	MT MTs	Microtubules microtubules microtubule
XMAP215		XMAP215
oscillations		oscillations
kinesin-1		kinesin-1 Kinesin-1
motor		motors motor Motors
microsphere		microsphere microspheres Microspheres
MCAK		MCAK
spindle		spindle
force		force forces

Table 3.5. Top ranked terms for Howard Group

Label	Abbrev.	Lexical variants
sphingolipid		sphingolipid sphingolipids Sphingolipid
lipid rafts		lipid rafts Lipid rafts
amyloid precursor protein	APP	amyloid precursor protein
Madin-Darby canine kidney	MDCK	Madin-Darby canine kidney
<i>protein</i>		protein Proteins proteins
membrane		membranes membrane
<i>cell</i>		cells cell
beta-secretase		beta-Secretase beta-secretase
lipid		lipids Lipids Lipid lipid
<i>raft</i>		rafts raft

Table 3.6. Top ranked terms for Simons Group

Label	Abbrev.	Lexical variants
force generators		force generators
pericentriolar material	PCM	pericentriolar material
oscillations		oscillations
RNA		RNA RNAs
SAS-4		SAS-4
centrioles		centrioles
pauses		pauses
spindle		spindle
force		force forces
mitotic spindle		mitotic spindle

Table 3.7. Top ranked terms for Grill Group

Label	Abbrev.	Lexical variants
neuroepithelial cell		neuroepithelial cell NE cells NE cell neuroepithelial cells
mouse		mouse Mouse
neurogenesis		Neurogenesis neurogenesis
prominin-1		prominin-1
neuroepithelial cell	NE	Neuroepithelial neuroepithelial cells cell
progenitor divisions		progenitors progenitor divisions
neuron		Neurons neuron neurons
prominin		prominins Prominin prominin

Table 3.8. Top ranked terms for Huttner Group

3.3 LMO Benchmark: *lipoprotein metabolism*

To assess the quality (Section 3.4) as well as the stability (Section 3.5) of the term ranking, as benchmark test set serves the manually defined domain ontology on lipoprotein metabolism which is rich in synonyms. All its terms are domain relevant.

The lipoprotein metabolism ontology The *Lipoprotein Metabolism Ontology* (LMO) was manually built in collaboration with domain experts from Unilever for the purpose of document retrieval. The LMO contains in total 522 concepts and 964 additional synonyms, with an average term length of 15 (2 words of 7.5 characters). The ontology contains for the analysis additional 302 concepts for nutrition terminology. A concept as used here consists of a concept label and optional synonymous terms. A term can be any word or phrase of relevance to the studied domain. Concerning the relations between the concepts, the mean number of parents is 2 (with a maximum of 3) and the mean number of siblings is 5 (with a maximum of 10).

PubMed queries for documents on lipoprotein metabolism During the development of the lipoprotein metabolism ontology (Alexopoulou et al., 2008) 14 terms were identified that retrieve domain relevant documents when submitted in a PubMed search. This process was performed empirically in collaboration with scientist from Unilever Research. This document set was now used as test set to reproduced the terminology contained in the manually created *Lipoprotein Metabolism Ontology*. The 14 terms are shown in Table 3.9 modified with restrictions to publication date 2006 or 2007 respectively.

3.4 Evaluation of the quality of generated terms

To measure and compare the quality of generated terms, two benchmark sets have been created. The LMO benchmark in the domain of lipoprotein metabolism resulted from the work creating the search engine *LMOPubMed* (Section 8.4). The ZEBET benchmark in the domain of animal testing alternatives resulted from the work

"cardiovascular disease" AND 2006[pdat]	obesity AND 2006[pdat]
"cardiovascular disease" AND 2007[pdat]	obesity AND 2007[pdat]
LDL AND 2006[pdat]	HDL AND 2006[pdat]
LDL AND 2007[pdat]	HDL AND 2007[pdat]
"Blood Pressure" AND 2006[pdat]	apolipoprotein AND 2006[pdat]
"Blood Pressure" AND 2007[pdat]	apolipoprotein AND 2007[pdat]
Dyslipidemias AND 2006[pdat]	fatty acid AND 2006[pdat]
Dyslipidemias AND 2007[pdat]	fatty acid AND 2007[pdat]
"Insulin Resistance" AND 2006[pdat]	cholesterol AND 2006[pdat]
"Insulin Resistance" AND 2007[pdat]	cholesterol AND 2007[pdat]
lipoprotein AND 2006[pdat]	high-density lipoprotein AND 2006[pdat]
lipoprotein AND 2007[pdat]	high-density lipoprotein AND 2007[pdat]
"lipoprotein metabolism" AND 2006[pdat]	triglyceride AND 2006[pdat]
"lipoprotein metabolism" AND 2007[pdat]	triglyceride AND 2007[pdat]

Table 3.9. Queries to PubMed used to generate lipoprotein specific terminology. The queries are used to retrieve scientific abstracts from PubMed which are the source for terms in domain of lipoprotein metabolism.

creating the *Go3R* search engine (Section 8.5). In a first experiment in Section 3.4.1 it was tested to what extend the notion of noun phrases from text as suitable ontology terms holds when compared to biomedical ontologies. In the second evaluation in Section 3.4.2 the term generation provided by the four systems TerMine, Text2Onto, OntoLearn and DOG4DAG has been evaluated against the LMO benchmark (Alexopoulou et al., 2008). The third and fourth experiments have been performed to find the best configuration for DOG4DAG regarding the ranking measure and the corpus used to capture domain relevance. Based on the ZEBET benchmark in Section 3.4.3, the C-Value measure as used in TerMine has been compared against the tf-idf measure as used in DOG4DAG. A final analysis in Section 3.4.4 was performed to compare pure frequency counts with the tf-idf measure and the probability of occurrence using PubMed or Google Web CT 5-grams as reference corpus.

3.4.1 Noun phrases as term candidates

Before analysing term generation based on benchmarks, we generally assess the likelihood of text-derived terms to be ontology terms. For this we generated terms by retrieving from PubMed abstracts for 1,000 randomly selected MeSH headings and evaluated whether the retrieved candidate terms overlap with existing ontologies, here GO, MeSH, any the OBO or UMLS ontologies.

Evaluation setup

We used the Gene Ontology (as of Nov 14, 2009) and 13 sub trees of MeSH2010 (as of Oct 29, 2009). We randomly selected 1,000 terms from MeSH for term generation. We generated terms from text which was obtained by searching PubMed for 250 scientific abstracts containing the MeSH term. During the evaluation terms have been automatically mapped to existing ontologies. For OBO we use the EBI Ontology Look-up Service, Dec 2009) and for the Unified Medical Language System (UMLS) version 2006AB. For term generation completeness is measured in terms of recall, the ration of retrieved relevant terms from all known relevant terms. Precision quantifies the portion from all generated terms which are indeed relevant. The F-measure is

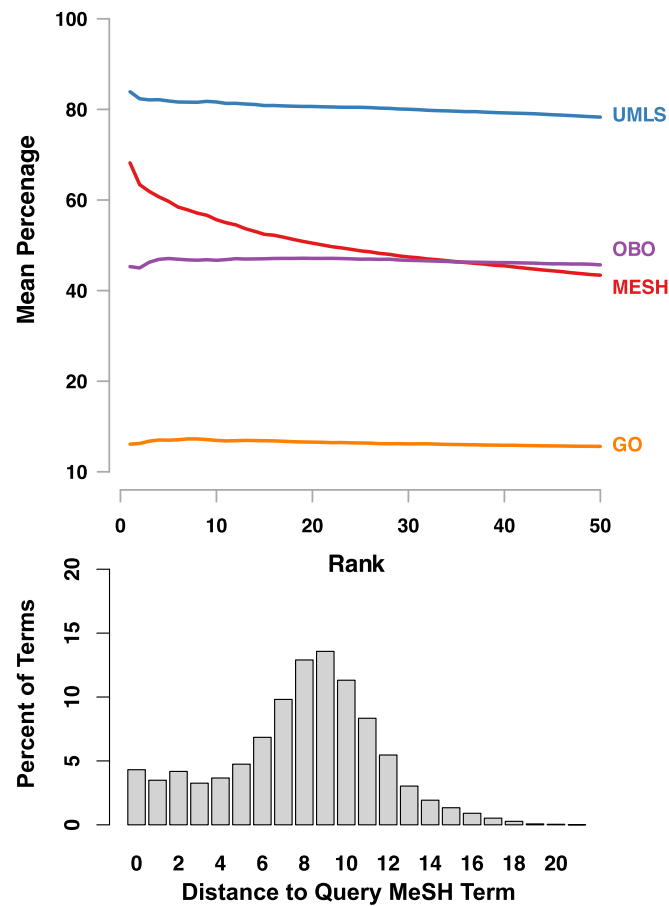


Fig. 3.3. Suitability of noun phrases as ontology term candidates. The mean percentage of generated terms from UMLS, MeSH, OBO, and GO in the top-k ranked generated terms and their distance to the randomly selected query MeSH term used to retrieve 250 PubMed abstracts. The generated terms show both, a high proportion of terms similar to existing ontology terms, justifying the notion of noun phrases as term candidates, and a certain variance of distances of generated terms to the query MeSH term, thus mapping out the neighbourhood of the query MeSH term as well as addressing other aspects of the document set.

the harmonic mean between precision and recall and allows to compare quality with respect to on numeric value.

Results

As the relevance of terms is subjective, we evaluated the quality by checking how many generated terms are already part of existing manually designed ontologies. This reveals whether significant noun-phrases according to our term generation have a similar structure to the manually defined terms. For 1,000 randomly selected terms from MeSH (Table 10.1) we generated terms on the basis of text from 250 PubMed abstracts per MeSH term and mapped the generated terms to GO, MeSH, OBO, and UMLS. Figure 3.3 (top) shows the mean percentage of generated terms which exist as term in these ontologies for the top-k ranked generated terms (for all generated terms and mapping to the ontologies see Table 10.2). Relatively independent of the number of terms, over 80% of the generated terms are similar to UMLS terms. Over 40% of the generated terms exist in OBO or MeSH. Results for UMLS are best since

Top	Precision				Average Precision			
	DOG4DAG	TerMine	Text2Onto	RelFreq	DOG4DAG	TerMine	Text2Onto	RelFreq
<i>LMO</i>								
50	35%	19%	17%	35%	65%	54%	38%	54%
200	20%	10%	12%	22%	42%	28%	23%	37%
1000	8%	4%	5%	8%	21%	12%	12%	20%
<i>LMO + Domain experts</i>								
50	75%	67%	33%	56%	86%	89%	52%	70%
200	55%	40%	49%	49%	74%	65%	38%	37%
1000	29%	20%	14%	28%	51%	40%	25%	45%

Table 3.10. Precision and Average Precision (rank dependent) for top 50, 200, and 1000 predictions. Four methods (DOG4DAG, Relative Frequency, TerMine, Text2Onto) are compared in terms of coverage of LMO and relevant vocabulary. The key finding is that among the top 1000 predictions there are up to 51% terms, which are in the LMO or considered good terms by expert, implying that automated term recognition can play an important role in semi-automated ontology design.

it is the largest terminology with nearly 10 million terms containing GO, MeSH and OBO. This shows that our notion of statistically significant noun phrases is a good approximation to manually defined term labels. The numbers for GO are lower (13%) since the GO terms usually do not appear literally in text. We analyzed the distance of the query MeSH term to the generated term if it exists in MeSH. Figure 3.3 (bottom) shows that 15% of terms map out the direct neighbourhood, i.e. are synonyms, siblings, parent, children, etc., having a distance to the query term ≤ 3 . Around 20% of terms are semantically distant and have a distance > 10 . Thus, the generated terms represent several possibly relevant aspects of the documents.

3.4.2 Comparison of different term generation methods

After showing that term generation in general is capable to retrieve suitable ontology terms from text we now evaluate different term generation methods using the validated relevant terms of the LMO benchmark. Five different ATR approaches, namely Text2Onto (Section 2.3.6), OntoLearn (Section 2.3.6), TerMine (Section 2.3.1) and two configurations of the method defined in this thesis DOG4DAG and RelFreq. DOG4DAG uses the tf-idf weight to rank terms, RelFreq used the relative frequency of terms in the corpus.

Evaluation setup

In Alexopoulou et al. (2008) we analysed the occurrence of Gene Ontology terms in PubMed abstracts to be able to define expectations for the task of ontology term generation. We found again, that less than 20% of Gene Ontology terms appear in PubMed abstracts and could hence be predicted. Out of the remaining terms, some could possibly be predicted because they have a definitional character, such as hydrolase acting on ester bonds, which comprises two noun phrases and a relation. For 53% of Gene Ontology terms, the contained noun phrases appear in a sentence in PubMed. Currently, no mature methods exist for finding such composite terms

Rank	DOG4DAG	RelFreq	TerMine	Text2Onto	OntoLearn
1	x metabolic syndrome	x review	x low-density lipoprotein	x patient	Mutation
2	x HDL	x metabolic syndrome	x cardiovascular disease	x disease	fish oil
3	x atherosclerosis	x diabetes	x metabolic syndrome	x risk	hypercholesterolaemia
4	x review	x atherosclerosis	x risk factor	x effect	Serum
5	x LDL	x HDL	x cardiovascular risk	x study	progression of atherosclerosis
6	x cardiovascular disease	x LDL	x high-density lipoprotein	x level	Apheresis
7	x diabetes	x cardiovascular disease	x low-density lipoprotein cholesterol	x atherosclerosis	omega-3
8	x dyslipidemia	x cholesterol	x high-density lipoprotein cholesterol	x cholesterol	treatment of hypertriglyceridemia
9	x high-density lipoprotein	x type	x fatty acid	x lipoprotein	reductase inhibitor
10	x cholesterol	x article	x coronary heart disease	x statin	Triglyceride
11	x low-density lipoprotein	x fatty acids	x coronary artery disease	x role	adhesion molecule
12	x cardiovascular risk	x high-density lipoprotein	x clinical trial	x syndrome	Evolution
13	x fatty acids	x role	x ldl cholesterol	x diabetes	purification process
14	x article	x dyslipidemia	x heart disease	x trial	Prescription omega-3
15	x insulin resistance	x low-density lipoprotein	x diabetes mellitus	x protein	omega-6
16	x type	x cardiovascular risk	x omega-3 fatty acid	x risk factor	hiv-infected
17	x statin	x hypertension	x blood pressure	x treatment	marker of inflammation
18	x hypertension	x combination	x oxidative stress	x event	strong evidence
19	x inflammation	x insulin resistance	x increased risk	x therapy	attractive target
20	x VLDL	x protein	x density lipoprotein	x review	accelerated atherosclerosis
21	x lipid metabolism	x disease	x cardiovascular risk factor	x type	internalization
22	x combination	x studies	x coronary artery	x mechanism	Scenario
23	x role	x inflammation	x statin therapy	x evidence	protease inhibitor
24	x oxidative stress	x association	x plant sterol	x development	inflammatory cell
25	x obesity	x plasma	x reverse cholesterol transport	x use	inflammatory marker

Table 3.11. The Top 25 lipoprotein related terms generated by four methods. The ranked candidate terms are shown for the methods DOG4DAG, RelFreq, TerMine, Text2Onto and OntoLearn. Terms have been predicted from 300 PubMed abstract retrieved for the query “lipoprotein metabolism” (limited to Review papers). Terms relevant to the lipoprotein metabolism domain are marked with x.

from text. For the Gene Ontology, methodologies are available for exploiting the structure of existing concepts to successfully propose new terms (Lee et al., 2006). Hence we evaluated the relevance of terms in ranked term lists as typical result of term generation methods. With the LMO benchmark a “lipoprotein metabolism”-specific corpus was created, consisting of 300 abstracts collected from PubMed with the query “lipoprotein metabolism” (limit for Review papers). These 300 abstracts were the maximal number of articles where all methods delivered results. Initially five different ATR methods should be tested on that corpus, namely Text2Onto (Section 2.3.6), OntoLearn (Section 2.3.6), TerMine (Section 2.3.1) and two configurations of the method defined in this thesis DOG4DAG and RelFreq. OntoLearn had to be excluded from further analysis, as it only generated a few terms so that a meaningful comparison would not be possible. Text2Onto was only included in the analysis for 300 abstracts as it was not possible to process all 3066 review article abstracts for “lipoprotein metabolism” listed in PubMed. We performed a bipartite analysis. We tried to automatically re-construct the manually created LMO terminology, compared the terms predicted by the four methods to the current LMO terms and also evaluated manually the top 1000 retrieved terms. All automatic comparisons between candidate terms and LMO were not case sensitive.

Results

For each of the four methods we list the top 25 ranked term (Table 3.11) and the percentage of relevant terms for the top 50, top 200, and top 1000 predictions. The results (Table 3.10) show that the precision for the top 50 predictions for LMO ranges from 17-35% and 4-8% for the top 1000 predictions. Using LMO and the expert terms leads to better results of up to 75% for the top 50 predictions and up to 29% for the top 1000. Considering the average precision and thus the ranking of terms, results for the top 50 predictions go up to 89% and for the top 1000 up to 51%.

	<i>LMO terminology predicted by DOG4DAG</i>		<i>LMO terminology literally contained</i>
	top 1000 generated terms	all generated terms	all contained in text
300 review abstracts for “lipoprotein metabolism”	8.75%	15.35%	20.98%
3,066 abstracts for “lipoprotein metabolism”	14.99%	38.25%	53.00%
50,000 abstracts containing “lipoprotein”			71.22%

Table 3.12. Coverage of LMO terminology in selected document sets. The coverage measures indicate the upper limit of terms that can be found with text-mining: Even a large text base with 50,000 documents contains only 71% of LMO terms. DOG4DAG can predict up to 38% of LMO terms.

Concerning recall (Table 3.12), 3066 documents contain only 53% of the LMO terms literally. DOG4DAG manages to predict up 39%, which is an encouraging result. Increasing the document base to 50,000 71% of the LMO terms are included indicating a possible upper limit of the percentage of ontology terms which can be generated on the basis of text.

3.4.3 Comparison of term ranking measures: C-Value vs. tf-idf

The C-Value method has been re-implemented and tested against the ZEBET benchmark to compare the two measures in the same system using the same linguistic component. Terms have been generated from all documents listed in the AnimAlt-ZEBET database, which constitutes a text corpus 305,344 words. The document frequencies required to calculate tf-idf were obtained from all PubMed abstracts.

ZEBET Benchmark: animal testing alternatives

This second benchmark has been created in collaboration with Dr. Barbara Grune (BfR/ZEBET) who manually curated the domain relevance for 3,271 terms generated from ZEBET database entries. The ZEBET database is the primary source for information on animal testing alternatives in Germany. It is mainly designed to support the examination of the imperative nature of animal experiments by providing information on possible alternative methods. In the ZEBET database only those methods are documented which fulfil at least one of the three following criteria:

- the method can be used to replace animal experiments (Replacement)
- the number of experimental animals is reduced (Reduction)
- the pain and suffering of the experimental animals are minimised (Refinement)

These criteria correspond to the scientific principle of the “3Rs” for the development of alternative methods to animal experiments which were developed by Russel and Burch and published in 1959 in their book “The Principles of Humane Experimental Techniques”. The documents in the ZEBET database are classified in data fields, e.g. designation of the method, keywords, evaluation, summary and references. The method summaries have been used as corpus for term generation.

(Description adapted from the ZEBET web site <http://www.bfr.bund.de/cd/1508>)

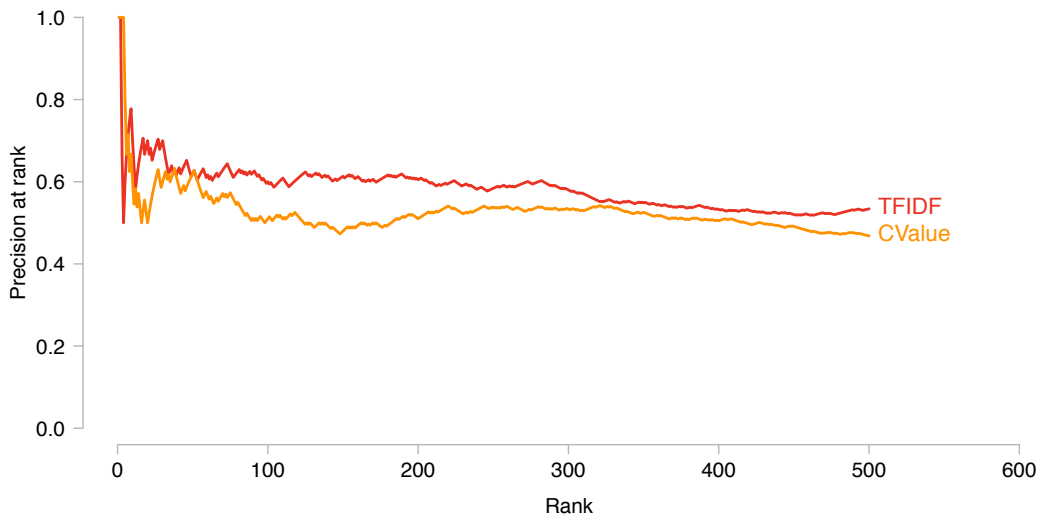


Fig. 3.4. Comparison of measures used in automatic term generation systems in the domain of animal testing alternatives: the C-Value measure vs. tf-idf. for the top 500 ranked terms. Both measures lead to similar results with more than 50% good terms. C-Value performs better in the top ranks and tf-idf performs better beginning from rank 10. The document frequency required to calculate tf-idf was obtained from PubMed abstracts.

Results

The results in Figure 3.4 show that both measures retrieve over 50% good terms. The C-Value method performs better in the top five terms while the tf-idf measure performs better in ranks below five, even though all single word terms are included in the ranked list. The C-Value does only retrieve multi-word terms which in English are likely to be technical term. Hence, the C-value measure has a-priori a lower likelihood to retrieve a general terms which are not domain relevant. C-Value prefers longer terms over shorter terms which is advantageous for the domain as the ZEBET database frequently mentions e.g. long names of animal test and alternative test methods. Examples are: “Draize rabbit eye test”, “eye irritation potential”, “mouse lethality assay”. Tf-idf ranks terms high which are frequently contained in the ZEBET documents and which are less frequent in the reference corpus relevant. This are e.g. “use” because “use of animals”, “assay”, or “test”, all are correct and domain relevant, but less informative. Overall tf-idf retrieves more domain relevant terms in the top 500 terms.

3.4.4 Quality of terms in dependence of the scoring method

Evaluation setup

With Alexopoulou et al. (2008) we evaluated several automatic term recognition methods and manually decided on the relevance of terms in the given domain of lipoprotein metabolism. In total 1,197 terms have been identified to be relevant in the LMO Benchmark (Section 3.3). In this experiment these terms are used to measure

the quality of the rankings obtained for several configurations of the term generation pipeline. Later, it will be experimentally investigated how much the background knowledge in form of a global reference corpus (Section 3.5.2) and the used model for scoring terms (Section 3.5.3) influence the results. Additionally it has been evaluated, whether the performance can be increased in meta analysis. For the meta analysis the mean probability of occurrence of a term in the local document set is used for ranking, where the probabilities are on one side calculated using the occurrence counts obtained from PubMed and on the other side those obtained from the Google n-gram source as explained in Section 6.3.2 (Global Frequency Revisions).

In the graphs in Figure 3.5 the performance of the different configurations of the term generation pipeline is shown by plotting rank-wise mean precision for the extraction of terminology on lipoprotein metabolism. Six methods have been defined:

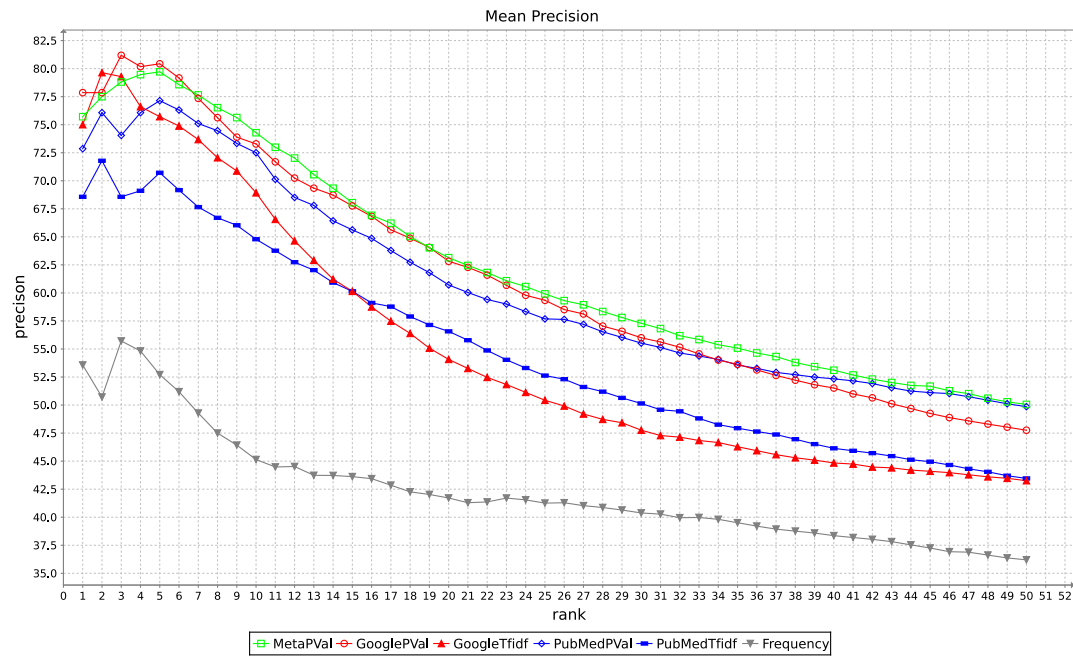
- *Frequency* – Term ranking with frequency of occurrence in the analysed document set.
- *PubMedTFIDF* – Term ranking with tf-idf, where the document frequency was derived from PubMed abstracts.
- *GoogleTFIDF* – Term ranking with tf-idf, where the document frequency was derived from the Google Web 1T 5-gram Version 1 (Brants and Franz, 2006).
- *PubMedPVALUE* – Term ranking with the probability of occurrence where the conditional probability is estimated with the probability of a terms occurrence in PubMed.
- *GooglePVALUE* – Term ranking with the probability of occurrence where the conditional probability is estimated with the probability of a terms occurrence according to the Google Web 1T 5-gram Version 1 (Brants and Franz, 2006).
- *MetaPVALUE* – Term ranking with the joint probability of *GooglePVALUE* and *PubMedPVALUE*.

For the graph in Figure 3.5(a) terms have been generated for 143 experiments. For 28 PubMed queries 50, 100, 500, 1000, and 2000 abstracts have been retrieved and terms have been generated. Additionally three manually created document sets related to the topic have been integrated in the analysis. To avoid a biased analysis the experiment has been repeated large scale by issuing queries for 811 concept labels in the LMO Ontology. The results have been shown in Figure 3.5(b).

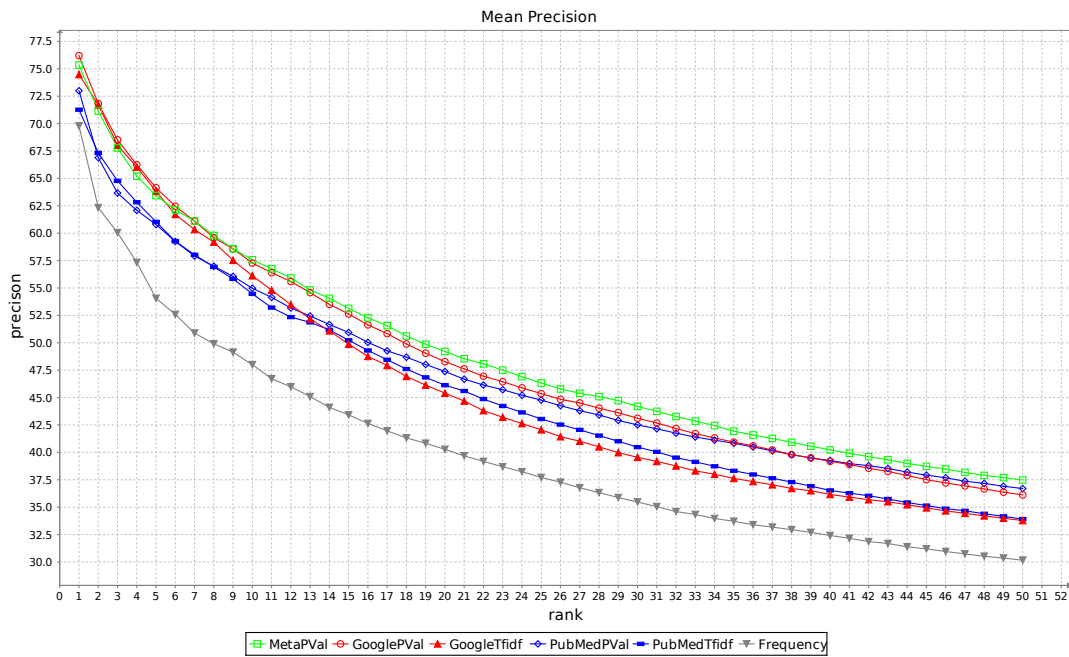
Results: Mean precision for generated terms from PubMed abstracts retrieved for 28 manually defined domain related PubMed queries

Details on the scoring and frequency assignment can be found Section 6.3.3 (Scoring Revisions) and Section 6.3.2 (Global Frequency Revisions).

For each rank the graph Figure 3.5(a) shows the overall percentage of predicted candidate terms which are contained in the 1,197 manually curated terms at this rank or further up in the hierarchy. For example a precision of 0.5 at rank 4 means that half (286) of the terms ranked 1-4 in the 143 experiments (in total 586) are known relevant terms in the domain. The following observations can be made:



(a) Terms generated based on text retrieved for 28 selected PubMed queries



(b) Terms generated based on text retrieved for PubMed queries for 811 existing LMO terms

Fig. 3.5. Quality of generated terms in the lipoprotein metabolism domain The mean precision is shown for the retrieval of terminology from lipoprotein metabolism related PubMed abstracts retrieved based on PubMed queries for (a) 28 selected representative terms (Section 3.3) and (b) for all 811 concept labels of lipoprotein metabolism ontology (LMO) terms. Only 13 terms standing for general concepts in the LMO (e.g human, long-lived person, or adolescent) have been excluded from the analysis. The retrieved terminology was compared against the manual created LMO.

- **Top 1:** For the top term can be seen that the precision is lower than for the top 2 or top 3. This is because for terms like “blood pressure” the term “pressure” or for “fatty acid” the term “acid” are the most frequently used terms but not contained in the ontology.
- **Top 3:** The graph shows that especially *GooglePVALUE* shows the most correct predictions with $> 77 - 81\%$ precision for the top ranks 1,2, and 3. *GoogleTFIDF* ($75 - 78\%$ precision) and *MetaPVALUE* ($76 - 78\%$ precision) show similar results. *PubMedPVALUE* follows just below with $73 - 76\%$ making more mistakes at rank 1 and 3. *PubMedTFIDF* achieves $68 - 72\%$ precision. All methods with background knowledge clearly outperform the *Frequency* method, which shows correct predictions for only $51 - 56\%$ of the terms.
- **Top 10:** *MetaPVALUE* performs best, followed by *GooglePVALUE*, and *PubMedPVALUE* all in the range of $72 - 74\%$ mean precision. *GoogleTFIDF* reaches 68% and *PubMedTFIDF* 65% mean precision. Also in the top 10 all methods using corpus statistics clearly outperform the *Frequency* method, which shows correct predictions for only 45% of the terms at rank 10.
- **Top 50:** *MetaPVALUE* and *PubMedPVALUE* perform best with on average 25 domain relevant terms within the top 50 terms (50%). *GooglePVALUE* follows with 48% , *PubMedTFIDF* and *GoogleTFIDF* with 43% , and *Frequency* with 36% .

Results: Mean precision for generated terms from PubMed abstracts retrieved for all LMO terms

In the previous experiment the 14 queries have been selected as queries likely to retrieve relevant terms. To ensure that the selection of the best performing configuration is not biased towards those queries, the experiment has been repeated for all but 13 labels of terms defined within LMO. The 13 terms with general meanings have been excluded because they are not specifically relevant for the domain “lipoprotein metabolism”. The terms middle-aged adult, middle-aged, hl, long-lived experimentee, long-lived person, long-lived population, young, young adult, enzyme, newborn, human, experimentee, and person have been excluded.

The obtained precision is lower, but the relative performance of the term generation configurations remains approximately the same.

Results: Pairwise comparison of chosen global corpus statistics and chosen statistical measure

Figure 3.6 compares the joint probability approach using PubMed and Google corpus statistics. Figure 3.7 shows the comparison between the tf-idf-based and probability-based methods. Figure 3.6(a) and Figure 3.6(b) show, that for this example the first 27 extracted terms were relevant terms. This indicates that once the document set contains terms of relevance the method ranks those high. It also shows that while *PubMedPVALUE* did fail to predict relevant terms at rank 18, 26 and 27 and *GooglePVALUE* at rank 21, 26, 27 the top 20 terms the combination *MetaPVALUE* lead in this case to a better ranking not showing this prediction of non-relevant terms.

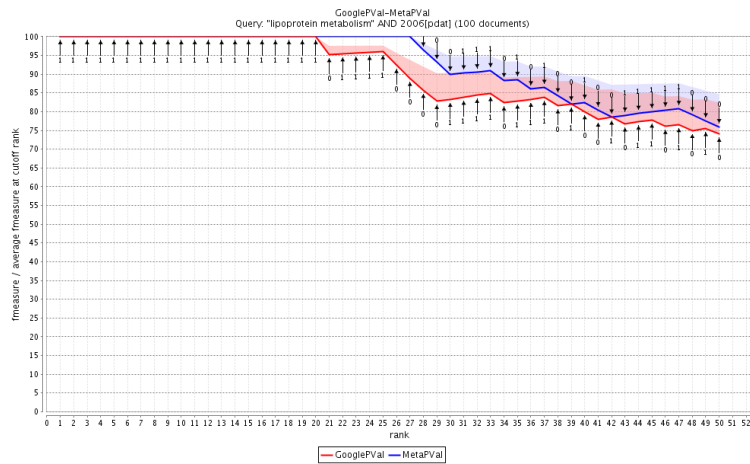
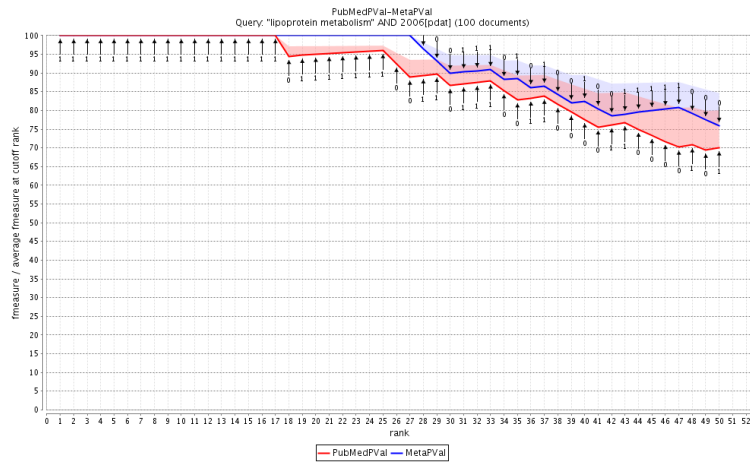
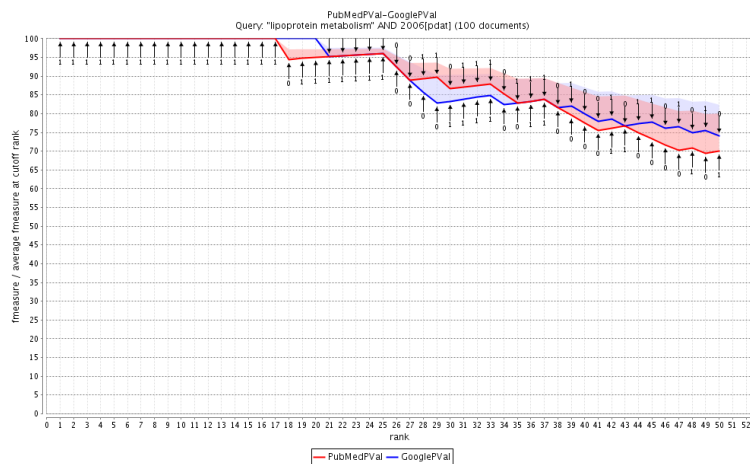
(a) Example where the *MetaPVALUE* outperforms *GooglePVALUE*(b) Example where the *MetaPVALUE* outperforms *PubMedPVALUE*(c) Comparison of *PubMedPVALUE* and *GooglePVALUE*

Fig. 3.6. Pairwise comparison of the quality of term generation in dependence of the reference corpus statistics. The F-measure (shaded: Average (rank-based) F-measure using average precision) is shown for ranges up to rank 50. Corpus statistics obtained from Google indexed web sites, PubMed abstracts or a Meta-approach combining both are compared. For this test set Google and PubMed corpus statistics perform equally well. The meta-approach combining both probabilities performs slightly better the single measures. Terms candidates have been generated from abstract retrieved via the PubMed query "lipoprotein metabolism" AND 2006[pdat]. The resulting terminology was compared against the manual created Lipoprotein Metabolism Ontology.

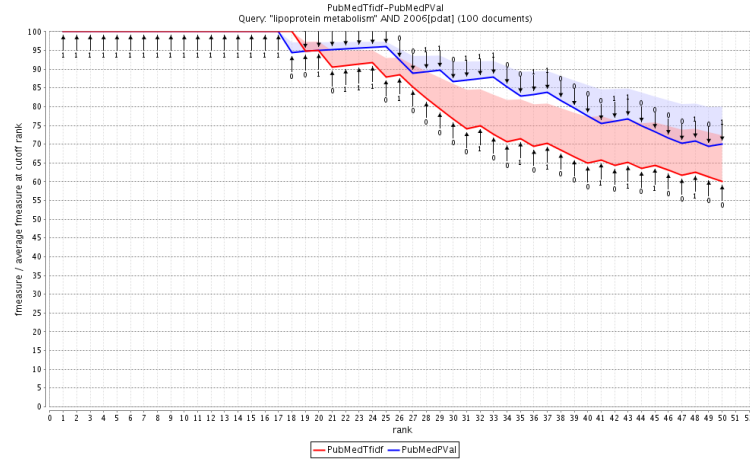
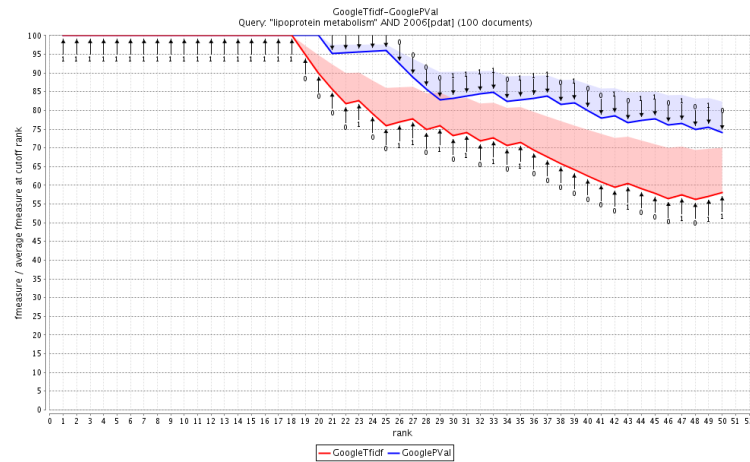
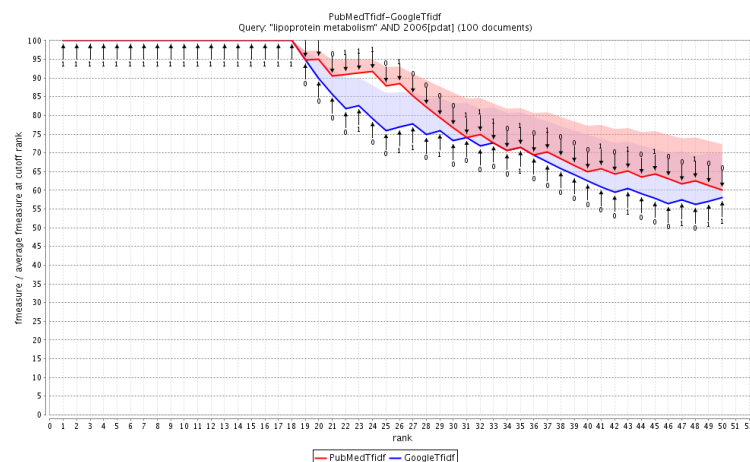
(a) Example where the *PubMedPVALUE* outperforms *PubMedTFIDF*(b) Example where the *GooglePVALUE* method outperforms *GoogleTFIDF*(c) Comparison of *PubMedTFIDF* and *GoogleTFIDF*

Fig. 3.7. Pairwise comparison of performance for term generation in dependence of the used statistical measure. The F-measure (shaded: Average (rank-based) F-measure using average precision) is shown for ranges up to rank 50. Relevance ranking using tf-idf is compared against the one using the true conditional probability. The less computational expensive method *TFIDF* is compared for the different corpus statistics. The approximation *TFIDF* performs lower than true probability (*PVALUE*). *TFIDF* performs better with PubMed than with Google corpus statistics. Terms candidates have been generated from abstracts retrieved via the PubMed query “lipoprotein metabolism” AND 2006[ptat]. The resulting terminology was compared against the manual created Lipoprotein Metabolism Ontology.

Summary of results

Scoring with the underlying probabilistic model performs better than the simplification tf-idf and both clearly outperform simple frequency measures on achieving a ranking of domain relevant vocabulary. The graphs in Figure 3.5 show that the differences between the probability-based methods are not significant. While *Google* corpus statistics lead to better terms in the top 5, *PubMed*-based corpus statistics were beneficial for the ranks from approximately 5 to rank 10. *GoogleTfidf* very good for the top 3 ranked terms, performs weaker in for terms ranked 4 to 50.

rank	background knowledge	statistical measure	method
5	Google n-grams	p-value	<i>GooglePVALUE</i>
10	Meta	p-value	<i>MetaPVALUE</i>
25	Meta	p-value	<i>MetaPVALUE</i>
50	Meta/PubMed	p-value	<i>MetaPVALUE/PubMedPVALUE</i>

Table 3.13. Summary of best global corpus statistics at different ranks. The best performing background knowledge and scoring method is shown for the best ranking for the top 5, top 10, top 25, and top 50 terms.

The method *GooglePVALUE* shows to be the best choice in terms of quality. In terms of runtime *PubMedTFIDF* is the best choice as the calculation of tf-idf is more efficient than the calculation of exact conditional probabilities (see Section 3.5.3 (Hypergeometric distribution)), and the corpus covers the required terminology but is significantly smaller. The PubMed corpus statistics are easier to handle and faster accessible. In the domain of lipoprotein metabolism the combination of the probabilities in the *MetaPVALUE* method do not lead to significantly better results.

The hypothesis “*Biomedical terminology including single word terms can be ranked better when weighting terms in contrast to large domain specific reference corpus.*” associated with research question 1 can be attested after the performed analysis. It has been surprisingly discovered that the ranking in contrast to a big enough general reference corpus leads to similar and better results than against domain specific reference corpus. This justifies in the biomedical domain the use of one huge source for stable word and phrase frequencies like the Google Web 1T 5-gram Version 1 (Brants and Franz, 2006) corpus and makes the methodology domain independent.

3.5 Stability of the ranking

The evaluation of term generation methods is difficult, as different opinions on the relevance or the quality of the terminology exist. For the task of ranking terminology according to its relevance in a domain it can be expected that results will differ significantly from subject to subject. The fact that the judgements on the quality of a ranking additional depend on the intended application makes the matter worse. Even if the assumption holds, that there does not exist the perfect ranking for terminology the dependencies of a ranking on technical or structural features

can be measured. It has been analysed how changes in the linguistic model (Part-Of-Speech, noun phrase detection) or the statistical model (scoring method, corpus statistic) influence the ranking of generated terms (Section 3.5).

- *quality of part of speech tags* (Section 3.5.1),
- *in-cooperation of background knowledge* (Section 3.5.2), or
- *underlying statistical measures* (Section 3.5.3).

Other parameters such as the *quantity of text* or the *selection of subsets of text* as input for term generation has been left out intentionally. It is clear that the predicted candidate terms highly depend on the text they have been extracted from. An analysis of the effects of varying input on the stability of the ranking is not useful, as no conclusions can be drawn from whatever results. Combining texts from e.g. two distinct domains will lead to a ranking as mixture of terms from both domains. Combining text from one domain with a random collection of other text certainly will not affect the ranking as much as the previous scenario. Practical experience showed, that the sequential processing of text collections from a specific domain lead to satisfactory term candidates; see Section 3.2 (Proof of concept) or Alexopoulou et al. (2008).

The following experiments were performed to evaluate the robustness of the obtained rankings. For several configurations of the term generation pipeline, the obtained terms lists were compared focusing on

- a terms change of rank,
- the percentage of terms contained in one but not the other term list, and
- the maximal change in rank observed for a majority of terms (90%).

These statistics correspond to the intended application scenario, where terms are selected from ranked list of terms. For this purpose, the stability of a ranking is dominated by the agreement between two rankings in the top segment of the ranked lists. A good agreement will show nearly no changes in rank for the most relevant terms, and will have only a moderate change in rank for less relevant terms. The less relevant a term is, the further down in the ranked list the term will appear. Assuming that the user is sequentially viewing the ranking, the likelihood that a term will be found by a user decreases the further down a term appears in the list. Hence the top terms (i.e. top 100) will be of most interest. Nonetheless, terms less relevant to the documents set still need to appear, as terms which are not contained in the ranked list cannot be found by searching or filtering. For the analysis on the quality of Part-Of-Speech tags the percentage of terms contained in one but not the other term list as well as the change in position is considered in the evaluation.

3.5.1 Dependency on the part-of-speech tagger

Hypothesis

The change of the implementation and model of the Part-Of-Speech tagger does not significantly change the extracted noun phrases or the the obtained rankings.

Experiment

For the 28 PubMed queries listed in Table 3.9 term rankings have been generated following the pipeline presented in Figure 3.1. The experiment has been repeated with two configurations using different Part-Of-Speech-taggers, namely the

- TNT part-of-speech tagger (Brants, 2000), trained on the Penn Treebank Corpus (English, Newspaper (Wall Street Journal), 1,200,000 tokens, 96.7% average accuracy, 0.13 standard deviation), and the
- LINGPIPE part-of-speech tagger (Carpenter, 2009), trained on the MedPost corpus (Smith et al., 2004).

Beside the pure ranking special interest in this experiment has been devoted to the number of terms extracted by with one but not the other configuration, as different Part-Of-Speech tags lead to different noun phrases and hence different candidate terms.

Results

Single examples: Figure 3.8 shows per example for the three domains *Blood Pressure*, *Obesity*, and *Insulin Resistance* how the ranking is influenced by a transition from one Part-Of-Speech tagging method to another one. The experiment was repeated with 50, 100, 500, and 2000 PubMed abstracts containing the words ‘Blood Pressure’, ‘Obesity’, or ‘Insulin Resistance’. The results for those examples indicate on one side that the ranking is relatively stable for those example. The distance of one terms rank in experiments with different Part-Of-Speech taggers is mostly below 2. The numeric results for the examples are listed in Table 3.14. While the change in rank did not seem to be of relevance in the top ranked terms, the listing shows that a significant number of terms either exist in one or another ranking, i.e. the number of missing terms in total is high (*not overlap (A)* or *not overlap (B)*).

The change of the Part-Of-Speech tagger does influence whether a term is predicted as candidate term at all.

Summary over 28 experiments The figures 3.9 and 3.10 show a summary plot over all 28 experiments for documents sets of the size 50, 100, 500, 1000, and 2000 PubMed abstracts. For each ranked term created using the TNT Part-Of-Speech tagger (**x-axis**) the plot illustrates the change in rank when exchanging the TNT Part-Of-Speech tagger with the LINGPIPE Part-Of-Speech tagger (**y-axis**). The differences in rank have been accumulated and are visualized using gray shadings in a hexagon plot. The darker a hexagon is displayed the more agreement was observed between the 28 experiments. The results are shown for the top 25 and top 100 ranked terms.

Missing terms Interpreting the summarised visualisation for the 28 experiments, it can be seen that in the top 25 ranked terms segment the number of missing terms decreased from 5.6% to 3.6% on average, when using more documents as basis for the term generation. For the top 100 ranked terms the proportion of missing terms decreases from 6.6% to 4.8%.

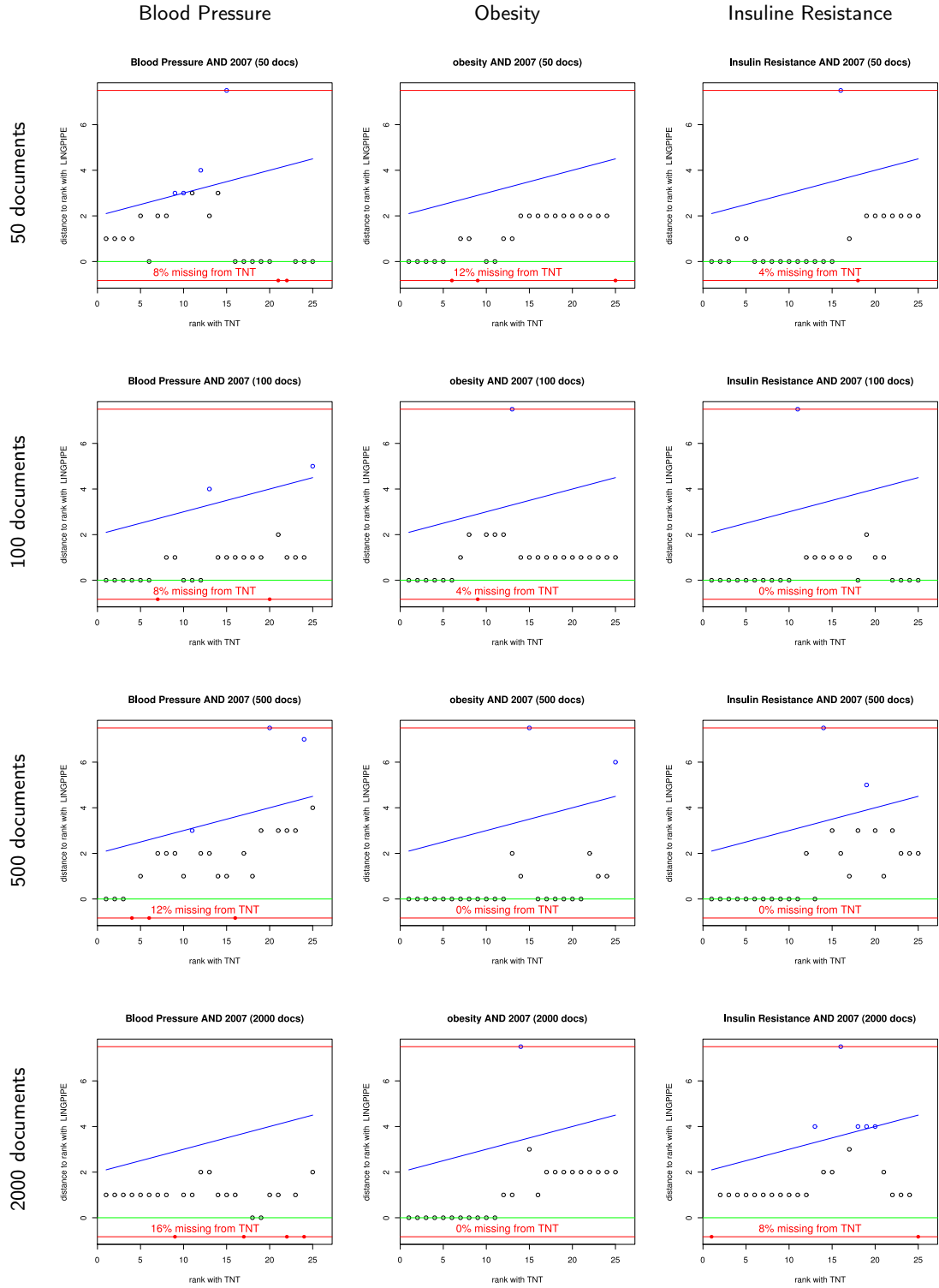


Fig. 3.8. Change in rank (selected samples): TNT vs. LingPipe Part-of-speech tagger.

Examples for rankings based on 50, 100, 500, and 2000 PubMed abstracts. The plot illustrates for each ranked term (x-axis) the change in rank when exchanging the TNT part-of-speech tagger with the LINGPIPE part-of-speech tagger (y-axis). Results are shown for the top 25 ranked terms. Terms missing in the ranking are plotted with negative distance in red. Terms which show a difference in ranks below $2 \pm (\text{rank} * 5\%)$ are plotted in black color, others in blue color. The threshold is illustrated by the gently inclined blue line.

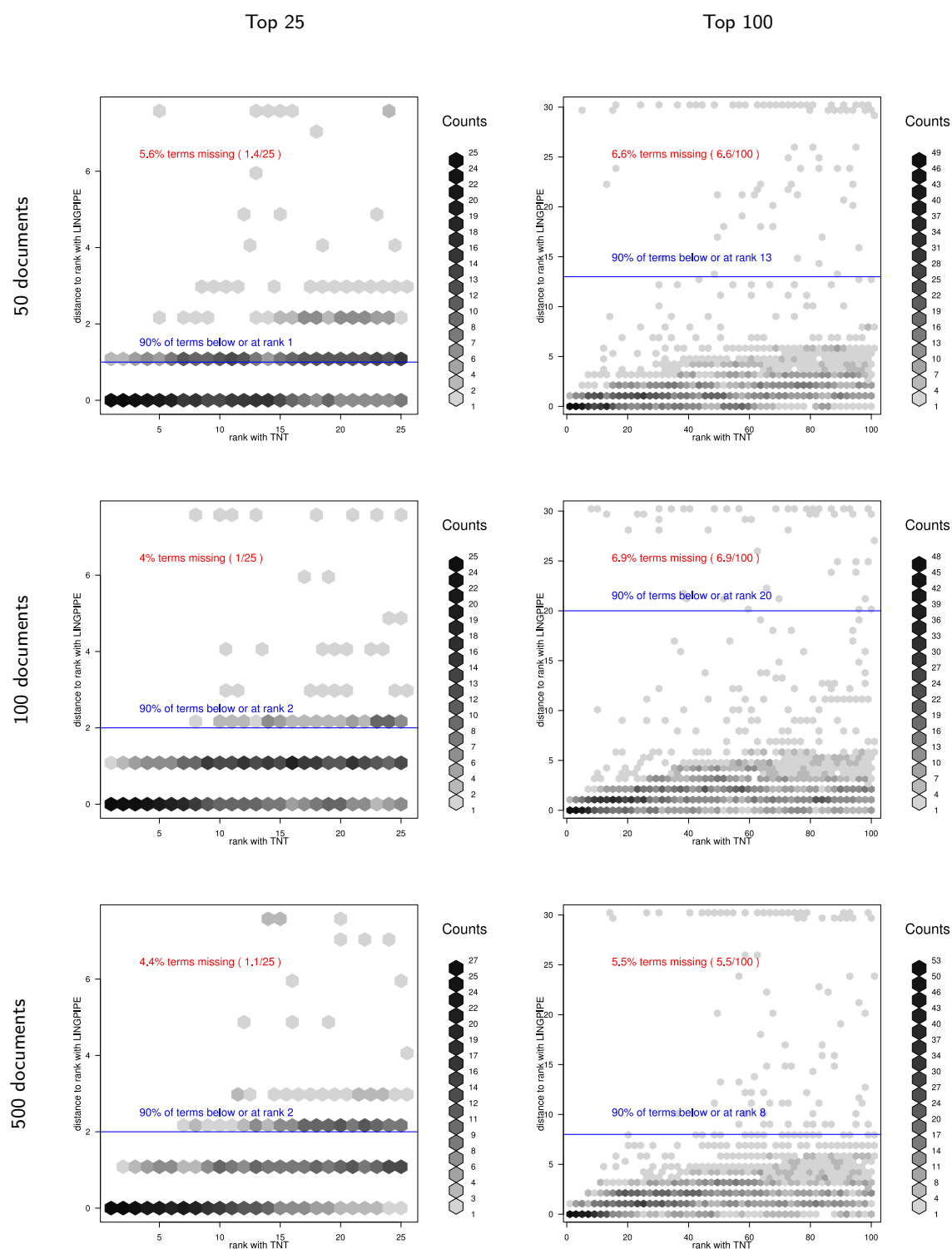


Fig. 3.9. Change in rank (summary): TNT vs. LINGPIPE Part-of-speech tagger (part 1)

Summary plot of term generation results on the basis of 50, 100, and 500 documents. The plot illustrates for each ranked term (x-axis) the change in rank (“below or at rank”) when exchanging the TNT part-of-speech tagger with the LINGPIPE part-of-speech tagger (y-axis). The experiment covers term generation results from all 28 PubMed queries given in `dataSetLipoProteinRelated`. The darker a hexagon is displayed the more agreement was observed throughout all experiments. Results are shown for the top 25 and top 100 ranked terms.

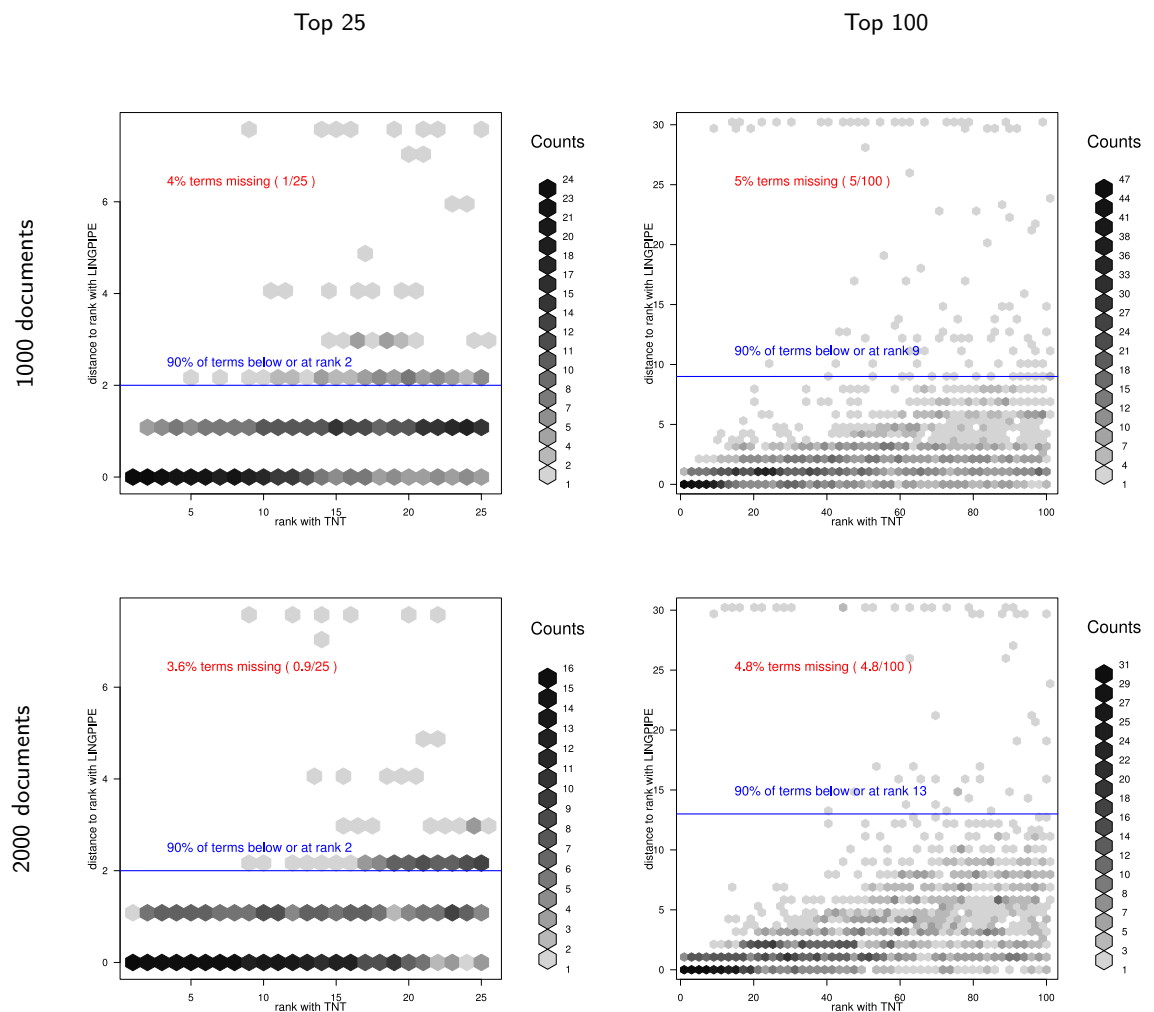


Fig. 3.10. Change in rank (summary): TNT vs. LINGPIPE Part-of-speech tagger (part 2)

Summary plot of term generation results on the basis of 1000 and 2000 documents. The plot illustrates for each ranked term (x-axis) the change in rank ("below or at rank") when exchanging the TNT part-of-speech tagger with the LINGPIPE part-of-speech tagger (y-axis). The experiment covers term generation results from all 28 PubMed queries given in `dataSetLipoProteinRelated`. The **darker** a hexagon is displayed the more agreement was observed throughout all experiments. Results are shown for the top 25 and top 100 ranked terms.

Difference in rank Independently from the number of documents 90% of terms show a shift in rank of below 2 when changing the Part-Of-Speech tagger (top 25). The more text is used for term generation (here test in the range from 50 to 2000 abstracts) the smaller is the average difference in rank. For the top 100 ranked terms the 90% borderline showed to be 13, 20, 8, 9, 13 for 50, 100, 500, 1000, and 2000 documents as basis for term generation.

The experiment suggests, that the agreement between the rankings is especially high in the top most segment. This can be explained as follows. Single word terms show higher frequency of occurrence in texts and lower dependency on the correct-

query string	result size	overlap	not over-lap (A)	not over-lap (B)	total (A)	total (B)	spear. correlation (A,B)
"Blood Pressure" AND 2007[pdat]	50	1992	489	466	2481	2458	0.97
"Insulin Resistance" AND 2007[pdat]	50	2034	478	477	2512	2511	0.95
obesity AND 2007[pdat]	50	1740	420	423	2160	2163	0.98
"Blood Pressure" AND 2007[pdat]	100	3618	895	841	4513	4459	0.97
"Insulin Resistance" AND 2007[pdat]	100	3646	876	838	4522	4484	0.96
obesity AND 2007[pdat]	100	3437	938	910	4375	4347	0.95
"Blood Pressure" AND 2007[pdat]	500	15310	4138	3776	19448	19084	0.96
"Insulin Resistance" AND 2007[pdat]	500	14705	4194	3814	18899	18519	0.95
obesity AND 2007[pdat]	500	14890	4307	4056	19197	18946	0.96
"Blood Pressure" AND 2007[pdat]	1000	28340	8294	7318	36634	35657	0.96
"Insulin Resistance" AND 2007[pdat]	1000	25686	7512	6695	33198	32383	0.95
obesity AND 2007[pdat]	1000	27189	8108	7444	35297	34632	0.96
"Blood Pressure" AND 2007[pdat]	2000	50529	15576	13428	66105	63957	0.95
"Insulin Resistance" AND 2007[pdat]	2000	45284	13906	12242	59190	57529	0.95
obesity AND 2007[pdat]	2000	46905	14229	12821	61134	59729	0.96

Table 3.14. Dependency of the term generation on the choice of the Part-of-speech tagger. Evaluation results comparing the rankings obtained for term generations experiment when changing the Part-of-Speech tagger for the example domains *Blood Pressure*, *Obesity*, *Insulin Resistance*. Over all experiments an average overlap of 79%(± 0.015) was observed.

ness of the assigned Part-Of-Speech tag. Single word terms will in general appear higher in the ranking than compound words. But the extraction of single word terms is also less vulnerable to variations in the assigned Part-Of-Speech tags as only one tag must follow the pattern for candidate terms (noun phrase pattern). The extraction of compound words on the other side requires the whole sequence of words to follow the noun phrase pattern.

3.5.2 Dependency on the global corpus: Google vs. PubMed

In this section it will be experimentally tested what influence the choice of the global corpus as basis for token and phrase frequencies has on the obtained rankings. Two sources for frequencies have been prepared to be used during term generation. For the Web 1T 5-grams Version 1 (Brants and Franz, 2006) with its 4.4 million 1,2,3,4, and 5-grams based on text of 1,024,908,267,229 tokens has been encapsulated as web service (**Google-N-Grams**). It contains n-grams from web sites indexed by Google in 2005. The sentence-based occurrence and co-occurrences counts of words in PubMed have been calculated and encapsulated as web service too (**PubMed-Cooc**).

Hypothesis

Reviewing the results presented in Section 3.4 (Evaluation of the quality of generated terms) it can be hypothesised that the influence of the background knowledge on the predicted candidate terms as well as on the ranking is evident. With a total difference in precision of 2.5 to 5% for the range from rank 10 to 50 the difference cannot be regarded as significant.

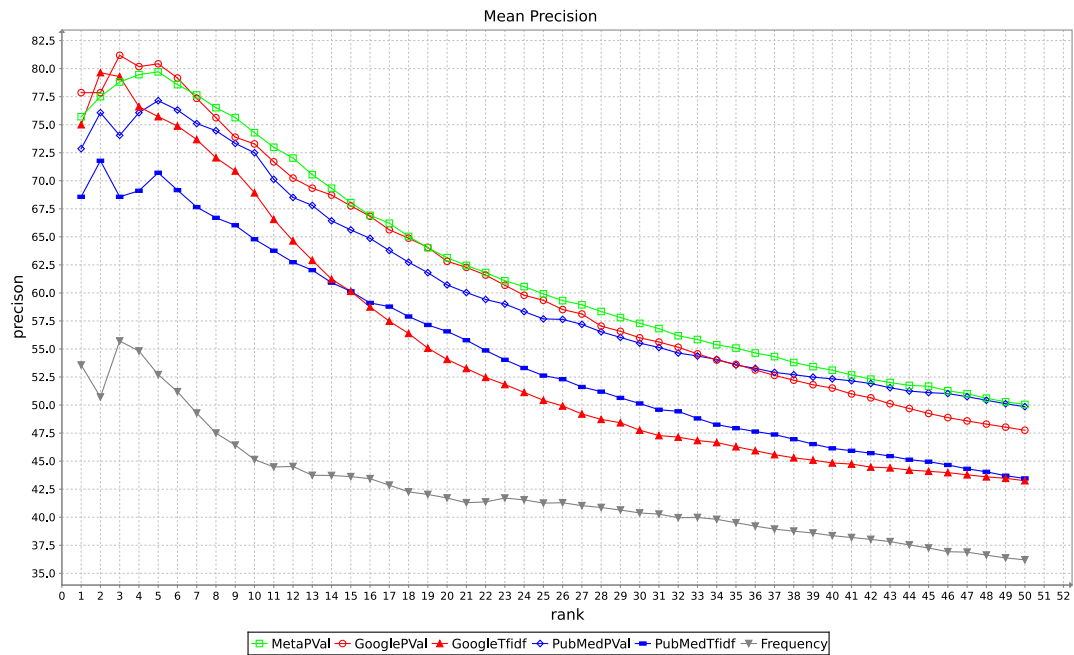


Fig. 3.11. Mean precision for the retrieval of terminology from lipoprotein metabolism domain, see Figure 3.5.

Experiment

For the 28 PubMed queries listed in Table 3.9 term rankings have been generated following the pipeline presented in Figure 3.1. The experiment has been repeated with two configurations using the different corpus statistics *Google – N – Grams* and *PubMed – Cooc* for scoring.

Beside the pure ranking special interest in this experiment has been devoted to the number of terms extracted with one but not the other configuration. Different Part-Of-Speech tags lead to different noun phrases and hence different candidate terms.

Results

Single examples Figure 3.14 shows per example for the three domains *Blood Pressure*, *Obesity*, and *Insulin Resistance* how the ranking is influenced when using instead of corpus statistics obtained from PubMed the statistics obtained from Google. The experiment was repeated with 50, 100, 500, and 2000 PubMed abstracts containing the words ‘Blood Pressure’, ‘Obesity’, or ‘Insulin Resistance’.

Summary over 28 experiments The figures Figure 3.12 and Figure 3.13 show a summary plot over all 28 experiments for documents sets of the size 50, 100, 500, 1000, and 2000 PubMed abstracts. For each ranked term which has been scored using occurrences and co-occurrences from PubMed (**x-axis**) the plot illustrates the change in rank when exchanging the PubMed occurrences and co-occurrences with the Google n-grams statistical information (**y-axis**) and vice versa.

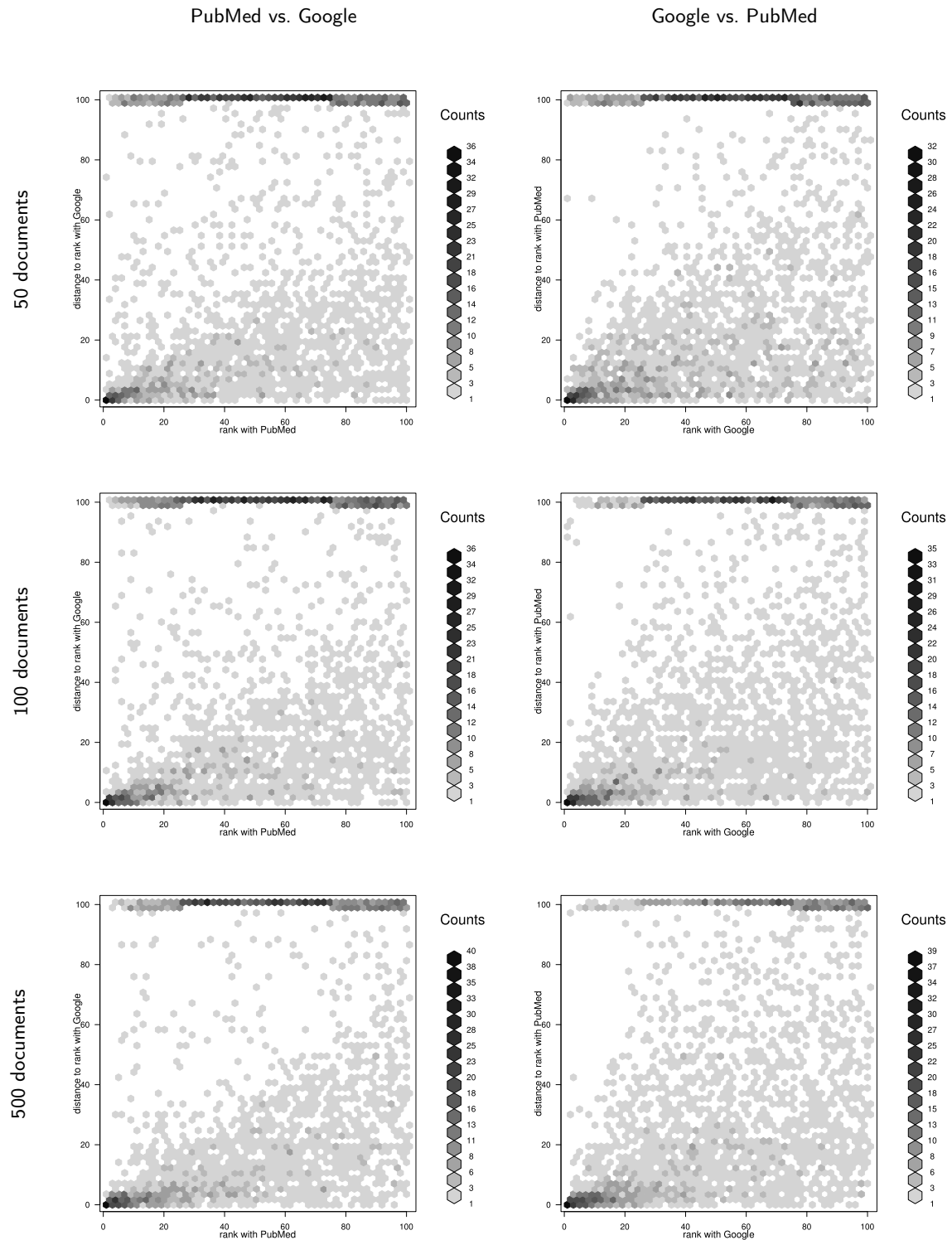


Fig. 3.12. Summary: PubMed vs. Google based corpus statistics (part 1)

Summary plot of term generation results on the basis of 50, 100, and 500 documents. The plot illustrates for each ranked term (x-axis) the change in rank when using instead of corpus statistics obtained from PubMed the statistics obtained from Google (y-axis) and vice versa. Results are shown for the top 25 ranked terms. The differences in rank have been accumulated over the 28 experiments and are visualized using gray shadings in a hexagon plot. The darker a hexagon is displayed the more agreement was observed between the 28 experiments. The results are shown for the top 100 ranked terms.

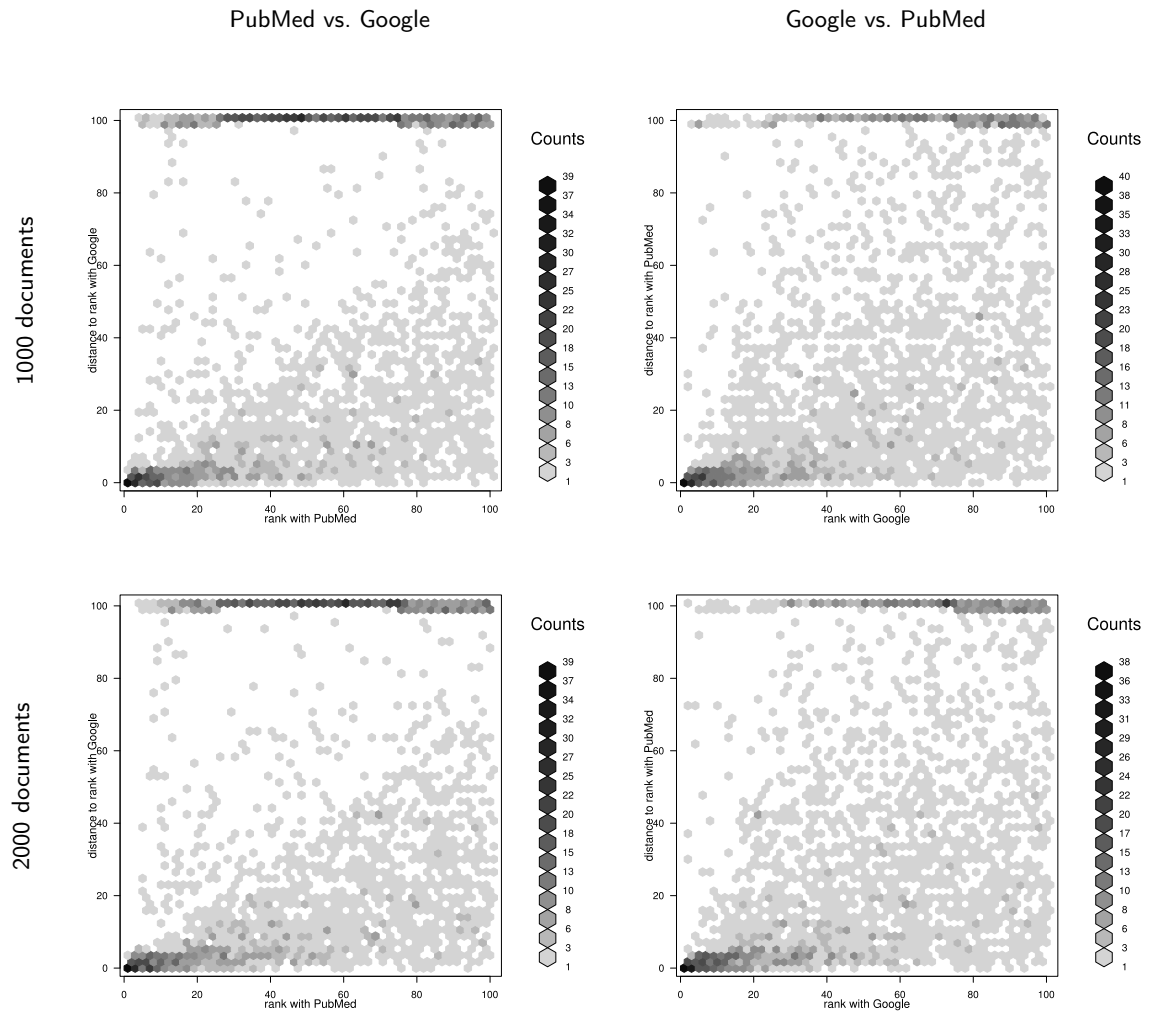


Fig. 3.13. Summary: PubMed vs. Google based corpus statistics (part 2)

Summary plot of term generation results on the basis of 1000 and 2000 documents. The plot illustrates for each ranked term (x-axis) the change in rank when using instead of corpus statistics obtained from PubMed the statistics obtained from Google (y-axis) and vice versa. Results are shown for the top 100 ranked terms. The differences in rank have been accumulated over the 28 experiments and are visualized using gray shadings in a hexagon plot. The darker a hexagon is displayed the more agreement was observed between the 28 experiments. The results are shown for the top100 ranked terms.

Missing terms As the difference in scoring does not affect the extraction of a candidate term, the extracted set of terms is identical for both configurations. This means no terms will be missed due to the change of the corpus statistics. From a applications point of view all terms can be found by searching and filtering.

Difference in rank The difference in rank observed in the plots is a contradiction to the hypothesis derived from the experiment in Section 3.4 (Evaluation of the quality of generated terms). The summary plots in Figure 3.12 and Figure 3.13 show change in rank greater than expected for a majority of candidate terms. The observation is independent from the direction of the comparison. The transition from

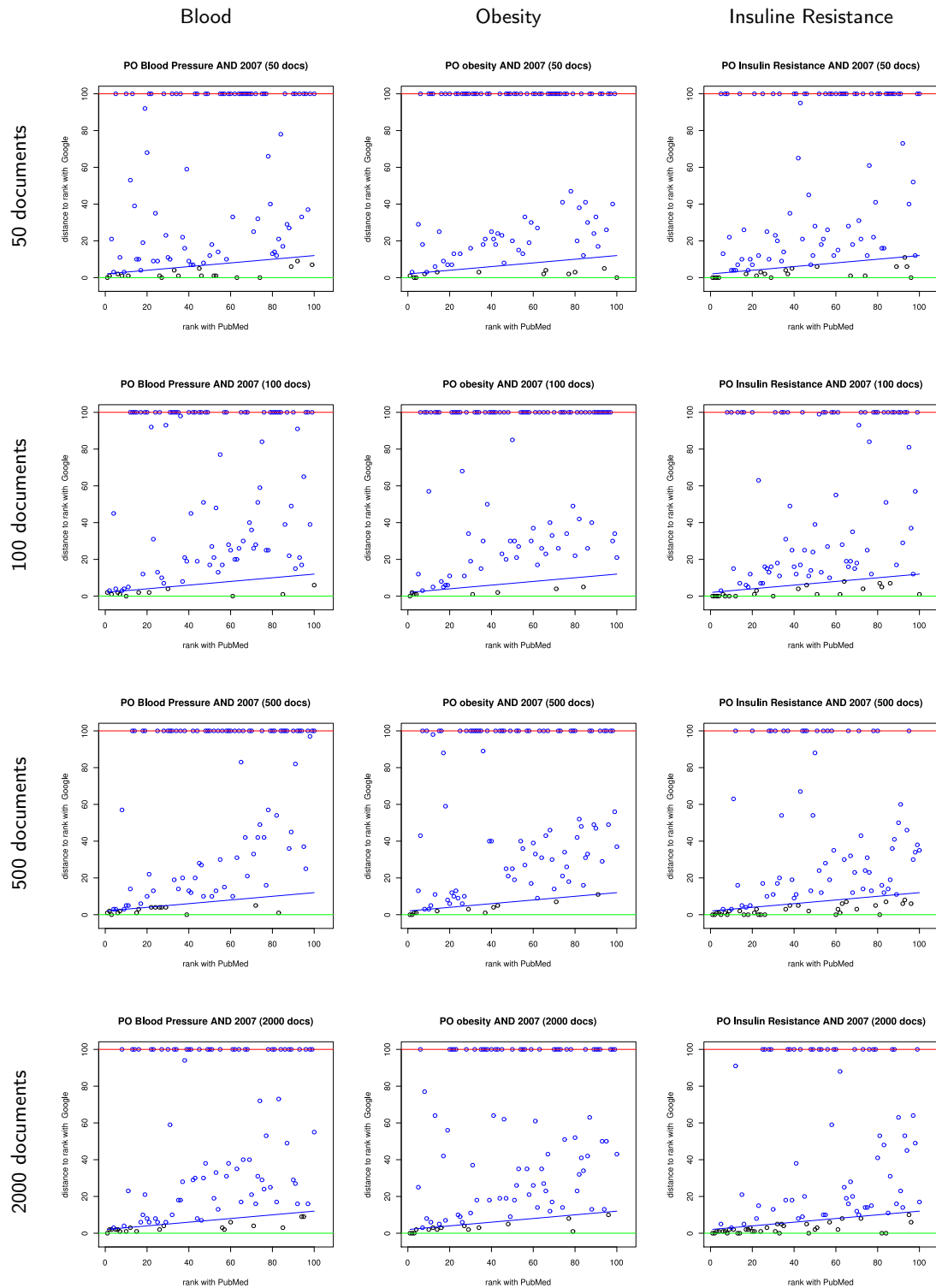


Fig. 3.14. Selected experiments: PubMed vs. Google based corpus statistics

Examples for rankings based on 50, 100, 500, and 2000 PubMed abstracts. The plot illustrates for each ranked term (x-axis) the change in rank when exchanging the available background knowledge obtained from PubMed with this obtained from Google (y-axis). Results are shown for the top 100 ranked terms. Terms missing in the ranking are plotted with negative distance in red. Terms which show a difference in ranks below $2 \pm (rank * 5\%)$ are plotted in black color, others in blue color. The threshold is illustrated by the gently inclined blue line.

(a) Top rank with PubMed-Cooc → lower rank with Google-N-Grams		
PubMed	Google	
8	134	[risk, Risk, risks, Risks]
19	203	[Waist, waist]
13	386	[Levels, level, levels]
12	1042	[Heart, hearts, heart]
15	9781	[index, Index]
15	9307	[Men, men]
17	9005	[WOMEN, Women, women]
15	11389	[Production, productions, production]
5	11503	[oils, Oils, oil]
16	12163	[fat, fats]
17	16450	[AI, A-I]
28	17065	[Alzheimer's disease, Alzheimer's Disease]
14	35395	[omega-3, Omega-3, omega3]
(b) Top rank with Google-N-Grams → lower ranked with PubMed-Cooc		
Google	PubMed	
19	53	[obese patients, obese patient]
19	59	[Serum, serum]
10	83	[Rats, Rat, rat, rats, RATS]
46	92	[protein, Proteins, proteins, Protein]
18	97	[lesions, lesion]
54	250	[concentration, concentrations, Concentrations]
9	427	[patients, patient, Patients]
29	10865	[adipokines, adipokine, Adipokines]
47	21094	[ApoA1, Apo-A1, apoA1, apoA-1, apo A-1, apo A1]
38	21159	[ApoCI, ApoC-I, apoC-I, apoCI, Apo C-I]
26	21327	[ApoA5, APOA5, apoA5]

Table 3.15. Examples for changes of the term ranking in dependence of the corpus statistics. Listing of concepts which are ranked significantly different when exchanging Google with PubMed corpus statistics and vice versa; (a) Google-N-Grams → PubMed-Cooc and (b) PubMed-Cooc → Google-N-Grams.

PubMed-Cooc to Google-N-Grams yields as many changes in rank as the transition from Google-N-Grams to PubMed-Cooc. The summary plots show that absolute difference in rank within the top 25 ranked terms is often below 6, but equally often above 25. Especially gene names or chemical compounds are prone to big difference in rank depending how many (replicated) web sites mentioning the gene are indexed in the web search engine. Table 3.15 lists a selection of those concepts where a the change of rank is especially high. The examples reach from

- commonly used terms like *man*, *woman*, *fat*, *risk* to
- terms not used in common language, e.g. *ApoA1*, *adipokines*, *ApoCI*, and
- terms with different meaning in non-biomedical text like *heart*, *level*, *production*, *AI*, and *omega-3*.

3.5.3 Dependency on the ranking score: tf-idf vs. probability of occurrence

In this section it will be shown to what extend the ranking varies when using tf-idf (term-frequency-inverse document frequency) weight as ranking score in comparison to the computationally more expensive computation of the conditional probability of occurrence of a candidate terms, from now on referred to as *TFIDF* and *PVALUE*.

Calculating probabilities The calculation of probabilities based on a hypergeometric distribution is computational expensive as the sum over all probabilities $\sum_0^{n_X} p(X)$ with n_X the number of occurrences of X has to be calculated.

HYPER-
GEOMETRIC
DISTRIBUTION

Hypergeometric distribution The hypergeometric distribution is a discrete probability distribution used in combinatorics. Assuming a finite number of elements, randomly n elements get drawn without replacement. The hypergeometric distribution describes the probability of drawing an element with the requested property.

Theorem 3.3. *The hypergeometric distribution depends on 3 parameters*

- the total number N of elements in the population
- the number $M \leq N$ of elements with a specific property contained in population
- the number $n \leq N$ of elements contained in the sample drawn

The probability distribution specifies the probability $P(X = k)$, that there exist k elements with the expected property in the sample.

$$h(k|N; M; n) := P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}},$$

with $\binom{N}{n}$ being the binomial coefficient.

The probability that there exist at most or at least k elements with the property in the sample is described with the cumulative sum of probabilities.

$$H(k|N; M; n) := P(X \leq k) = \sum_{y=0}^k h(y|N; M; n) = \sum_{y=0}^k \frac{\binom{M}{y} \binom{N-M}{n-y}}{\binom{N}{n}},$$

The hypergeometric distribution can be approximated for experiments with a small sample size and with very big total number of elements in the population. If $N \gg n$ applies, the hypergeometric distribution can be approximated by the binomial distribution without a significant error (Mosler and Schmid, 2006) (approximately $\frac{n}{N} \leq 0.05$).

Hypothesis

The estimation of the true probability of occurrence leads to more stable results than the simplification tf-idf, which is frequently used in term ranking.

Experiment

Compare the different rankings using the same global knowledge and discuss the cases where tf-idf does not rank similar to the probabilities.

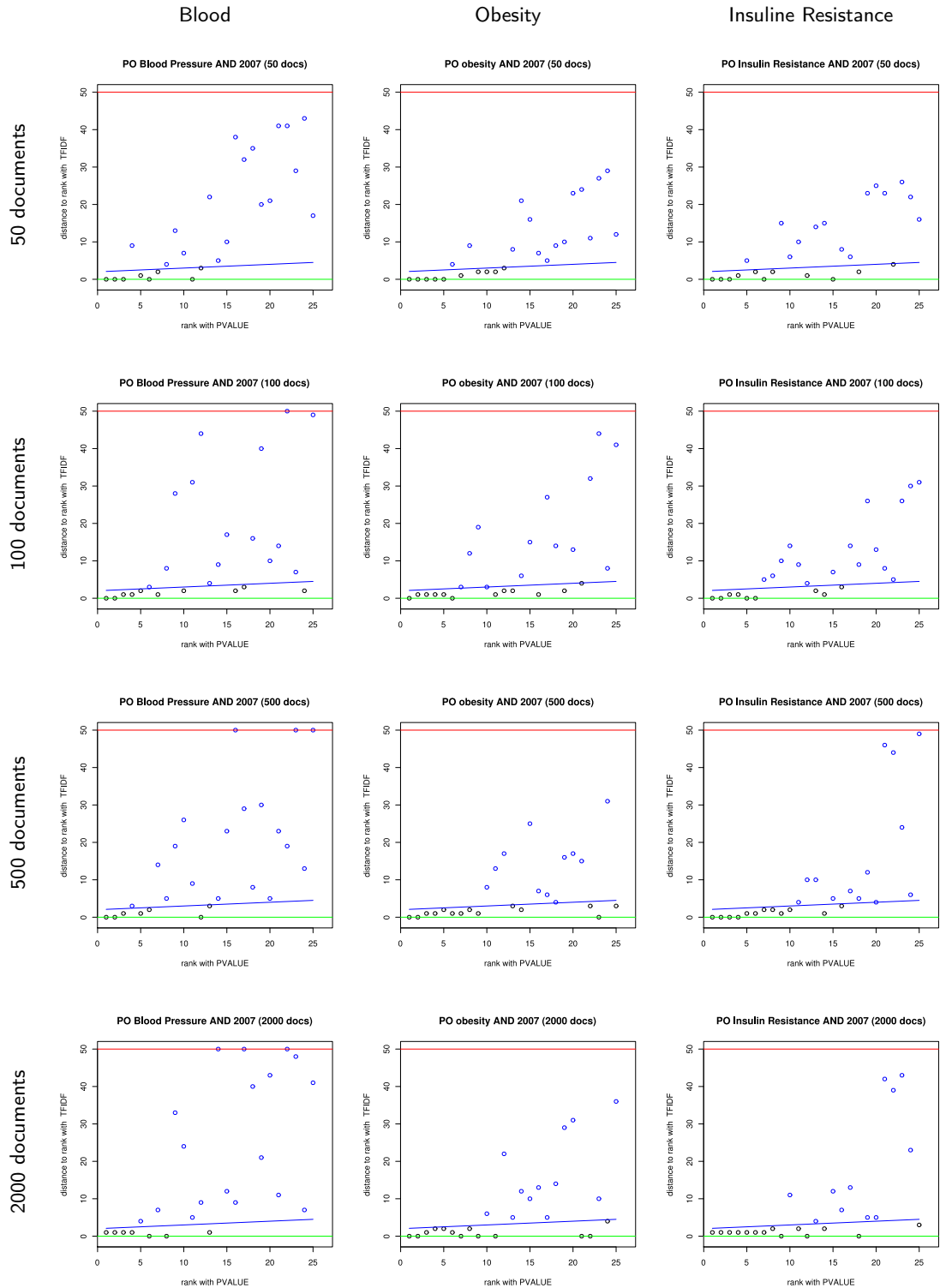


Fig. 3.15. Selected experiments: probability of occurrence (PVALUE) vs. TF-IDF

Examples for rankings based on 50, 100, 500, and 2000 PubMed abstracts. The plot illustrates for each ranked term (*x-axis*) the change in rank when calculating the relevance score instead of using true probabilities on the basis of the simplification TFIDF (*y-axis*). Results are shown for the top 25 ranked terms. Terms missing in the ranking are plotted with negative distance in red. Terms which show a difference in ranks below $2 \pm (rank * 5\%)$ are plotted in black color, others in blue color. The threshold is illustrated by the gently inclined blue line.

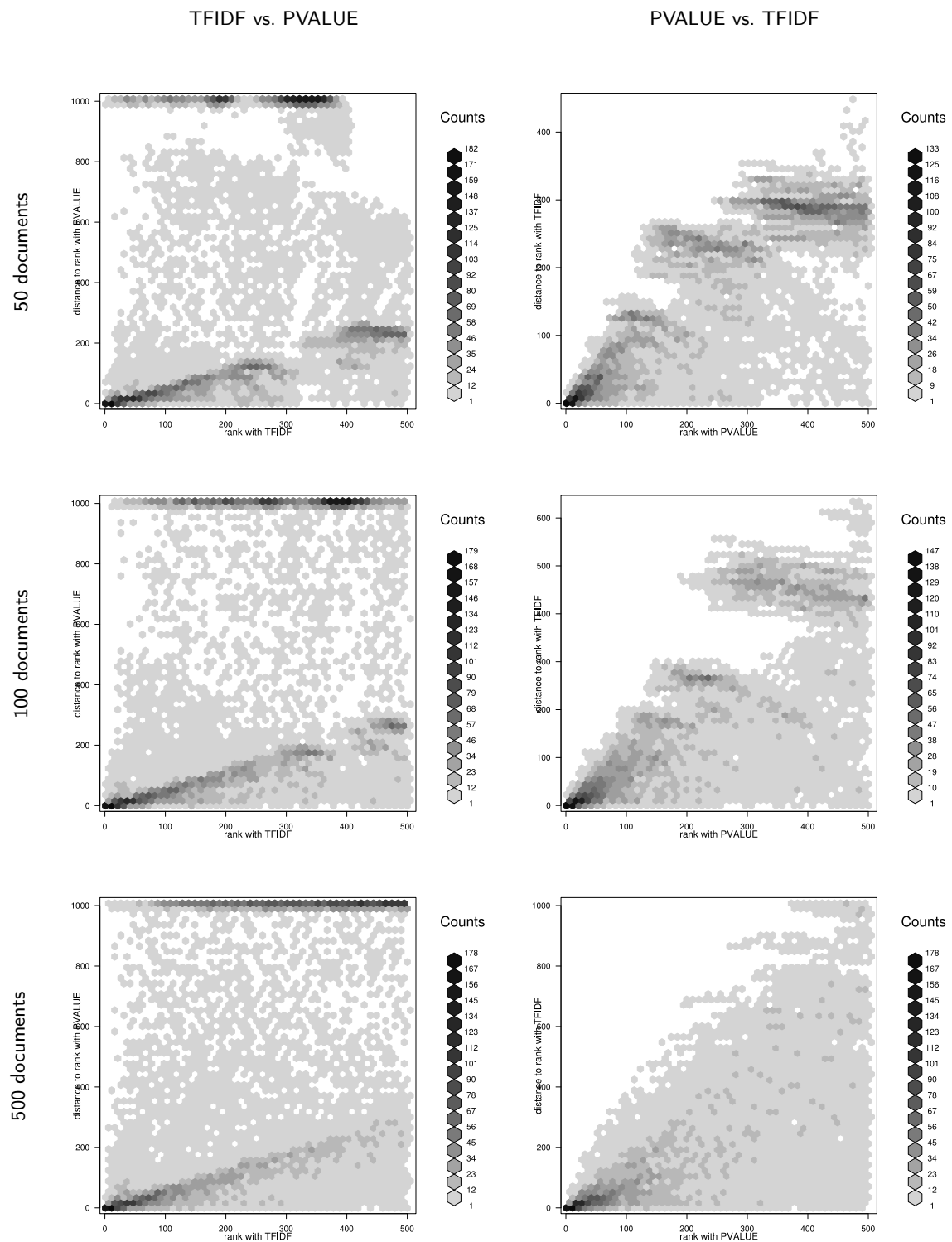


Fig. 3.16. Summary: TFIDF vs. probability of occurrence (PVALUE) (part 1)

Summary plot of term generation results on the basis of 50, 100, and 500 documents. The plot illustrates for each ranked term (x-axis) the change in rank when calculating the relevance score instead of using TFIDF on the basis of true probabilities (y-axis) and vice versa. The experiment covers term generation results from all 28 PubMed queries given in Table 3.9. The **darker** a hexagon is displayed the more agreement was observed throughout all experiments. Results shown for the top 500 ranked terms that many top ranked terms with TFIDF are not in the top with PVALUE, while top ranked terms with PVALUE are ranked high with TFIDF.

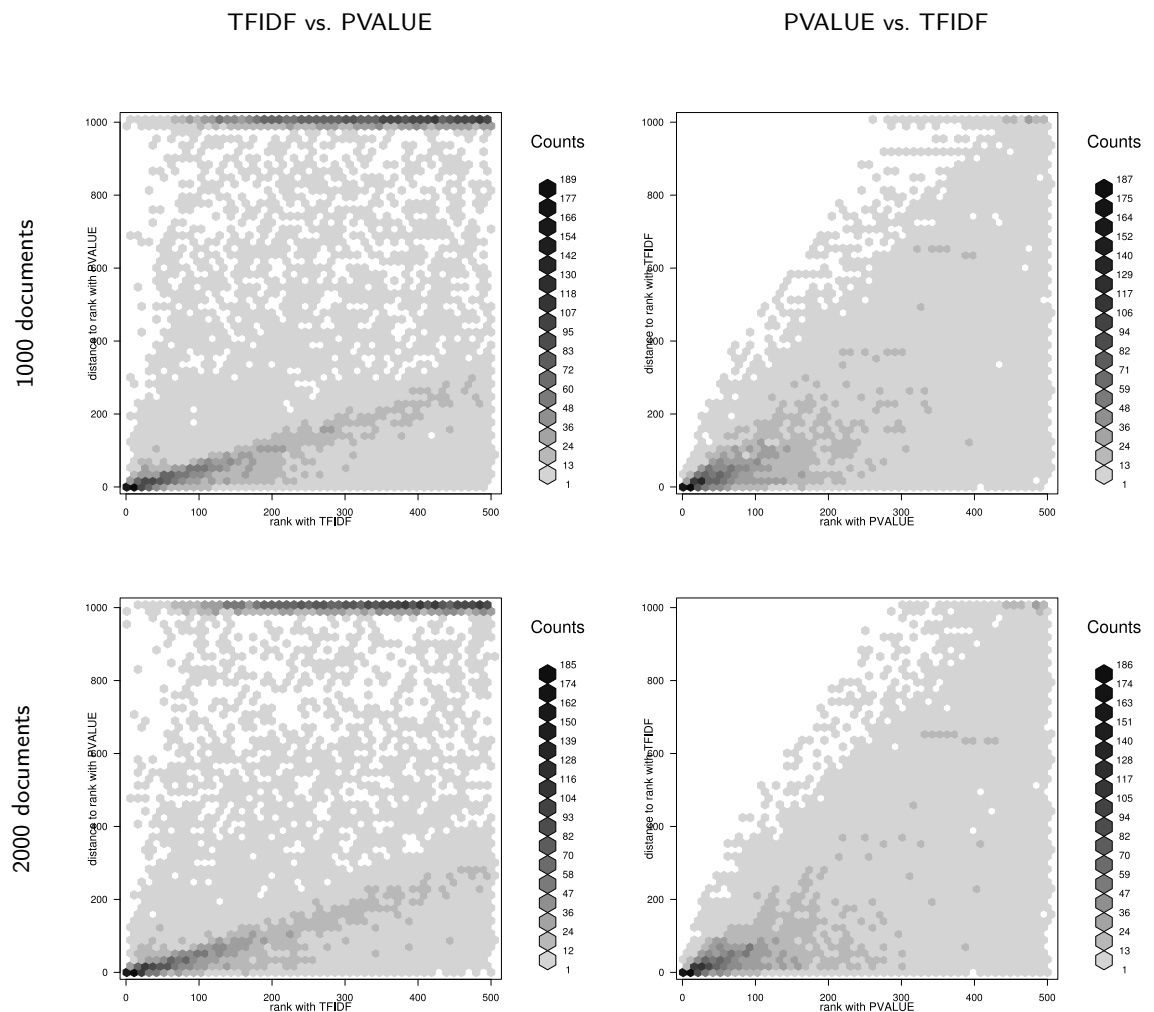


Fig. 3.17. Summary: TFIDF vs. probability of occurrence (PVALUE) (part 2)

Summary plot of term generation results on the basis of 1000 and 2000 documents. The plot illustrates for each ranked term (**x-axis**) the change in rank when calculating the relevance score instead of using TFIDF on the basis of true probabilities (**y-axis**) and vice versa. The experiment covers term generation results from all 28 PubMed queries given in Table 3.9. The **darker** a hexagon is displayed the more agreement was observed throughout all experiments. Results shown for the top 500 ranked terms that many top ranked terms with TFIDF are not in the top with PVALUE, while top ranked terms with PVALUE are ranked high with TFIDF.

Results

The results for this experiment look similar to those in Section 3.5.2 (Dependency on the global corpus: Google vs. PubMed).

Summary over 28 experiments The figures Figure 3.16 and Figure 3.17 show a summary plot over all 28 experiments for documents sets of the size 50, 100, 500, 1000, and 2000 PubMed abstracts. For each ranked term which has been scored using tf-idf and occurrences and co-occurrences from PubMed (**x-axis**) the plot illustrates

the change in rank when exchanging tf-idf with the probability of occurrences within PubMed documents.

Single examples Figure 3.15 shows per example for the three domains *Blood Pressure*, *Obesity*, and *Insulin Resistance* how the ranking is influenced when using instead of the simplification tf-idf a probabilistic measure assuming a hypergeometric distribution. The experiment was repeated with 50, 100, 500, and 2000 PubMed abstracts containing the words ‘Blood Pressure’, ‘Obesity’, or ‘Insulin Resistance’.

Missing terms As the difference in scoring does not affect the extraction of a candidate term, the extracted set of terms is identical for both configurations. This means no terms will be missed due to the change of the corpus statistics. From an applications point of view all terms can be found by searching and filtering.

Difference in rank As the experiment in Section 3.4 (Evaluation of the quality of generated terms) already suggested, also the summary plots in Figure 3.16 and Figure 3.17 show a change in rank for a majority of candidate terms. The ranking of terms in the top segment of the ranking is more stable than for lower ranks. Examples on single experiments for *Blood Pressure*, *Obesity*, and *Insulin Resistance* (Figure 3.15) show likewise, that the top 5 terms are extracted no matter which method is used. The same can be observed in the summary plots as well as in Figure 3.5 showing the qualitative comparison of the proposed method configurations. The observations in general are dependent from the direction of the comparison. There are more top ranked terms with *TFIDF* which change in rank in the order of magnitude compared to a ranking with the *PVALUE*-method. Terms ranked *PVALUE* also show changes in rank compared to a ranking with the *TFIDF*, but the differences are significantly lower. It is also evident, that the larger amount of text, i.e. a higher number of documents, increase the similarity between the rankings.

3.6 Summary and Discussion

An efficient term generation method has been designed, implemented, and evaluated as part of the *Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG)*. The method is capable of retrieving high quality domain relevant terms also found in manual created ontologies. An initial experiment confirmed that up to 80% of the generated terms already exist in the UMLS, which is with nearly 10 million terms the largest resources for terminology in the life sciences. The existing terms should be re-used or referenced (Section 3.4.1). In general this analysis shows that the notion of statistically significant noun phrases is a good approximation to manually defined term labels.

For the in-depth analysis two benchmarks have been defined. As benchmark for the comparative evaluation of several methods served an ontology containing 522 concepts with 964 synonyms which has been created in collaboration with Unilever Research UK in the domain of lipoprotein metabolism. The generated ranked lists of terms have additionally be validated by domain experts.

In the second domain of “Animal testing alternatives” an expert judged on the domain relevance for 3,271 terms.

The stability of the ranking has been analysed in detail and here the dependency on the selected Part-Of-Speech tagger for noun phrase extraction, the source for the global corpus statistics, as well as the used statistical model for term ranking.

Comparison of different term generation methods

On the basis of the lipoprotein metabolism benchmark it has been tested how each methods can recover the terminology contained in a manually defined ontology such as the lipoprotein metabolism ontology. *DOG4DAG* and three other tools that provide term recognition functionality, namely Text2Onto, OntoLearn, and TerMine have been used out-of-the-box.

The methods were capable of extracting terms of which 17 – 35% also existed in the ontology. The manual evaluation of term lists generated by the methods showed that 33 – 75% were good candidate terms. Later, these terms have been added to the ontology. *DOG4DAG* performed best in this evaluation. TerMine achieves with 89% average precision the best ranking in the top 50 retrieved terms. Average precision is a retrieval order dependent precision measure.

Coverage evaluated against the manually created lipoprotein metabolism ontology (LMO) was less then 15% (of possible 21%) when considering the terms retrieved based on 300 review papers and less then 38% (of possible 53%) when generating terms from 3,066 scientific abstracts containing the phrase “lipoprotein metabolism” in PubMed. This shows that by far not all terms in a the manually defined ontology can be found in scientific abstracts and that the selection of the text document is important.

The analysis showed that TerMine is capable of retrieving relevant terms. It retrieves less relevant terms, because it does not consider single words as term candidates. For a better ranking performance of Text2Onto, the mechanism of creating specialised rules provided by the Text2Onto framework should be used.

Qualitative analysis - lipoprotein metabolism

In a similar experiment the previous analysis was extended to find the best configuration of the term generation pipeline (Section 3.1). Now documents have not been manually selected, but instead there have been PubMed abstracts retrieved for 28 potentially relevant PubMed queries (Table 3.9).

The results in Section 3.4.4 (Quality of terms in dependence of the scoring method) show that terminology can be predicted with reasonable precision of 60% and above within the top 25 ranked terms. The best configuration reaches a precision of more than 80% (top 3), 74% (top 10), 60% (top 25), and 50% (top 50). According to this evaluation, 25 out of 50 predicted terms were relevant to the domain.

The same experiment has been repeated for all term labels from the Lipoprotein Metabolism Ontology. This way it was simulating that a user could start with any term to extend the ontology by generating new terms and still would be able to generate terms with a good precision. Again PubMed abstracts have been retrieved for each term label and terms have been generated with all configurations of the term generation pipeline.

Both analyses showed that *PVALUE*, which uses the true probability of occurrence, leads to better results than *TFIDF*.

It has been discovered that the ranking in contrast to a big enough general reference corpus leads to similar and better results ranking in contrast to the domain specific reference corpus. This suggests, that in the biomedical domain the same huge source for stable word and phrase frequencies as provided by the Google Web 1T 5-gram Version 1 (Brants and Franz, 2006) corpus can be used as in other domains. As no frequency information specifically from the domain is required the term generation in DOG4DAG can be regarded as domain independent.

An additional experiment to compare performance independently from the linguistic filter used to extract term candidates revealed that both term weighting with tf-idf as used in DOG4DAG and the C-Value method as used in TerMine were suitable to retrieve over 50% relevant in the domain of alternatives to animal testing.

The hypothesis that “Automatic term recognition methods can be extended to include single word terms while sustaining the high quality of the retrieval of domain relevant terminology.” associated with research question 1 can be attested after the performed analysis.

Stability of the ranking

In Section 3.5 (Stability of the ranking) the effects on the selection of the *quality of part of speech tags* (Section 3.5.1), the *in-cooperation of background knowledge* (Section 3.5.2), and the *underlying statistical measures* (Section 3.5.3) on the stability of the term candidate ranking have been tested for the proposed term generation pipeline.

- Part-Of-Speech tags have shown to have a low influence on the order of terms in the ranking, but showed to influence the extraction of a term in general. In Section 3.5.1 (Dependency on the part-of-speech tagger) it was experimentally shown that 5 to 7% of terms contained in the top 100 predicted candidate terms will not be extracted from text when changing the Part-Of-Speech tagger.

- The experiment Section 3.5.2 (Dependency on the global corpus: Google vs. PubMed) was performed to answer the question on the influence of the reference corpus to obtain occurrence counts for terms on the final ranking. The occurrence counts only influence the order of the terms in the ranking not the extraction itself. The experiment shows significant changes in the rankings depending on the corpus. The ranking is relatively stable in the top 25 ranked terms. Even though the ranking underlies high variations for some terms. The experiment Section 3.4.4 (Quality of terms in dependence of the scoring method) shows, that the qualitative difference is less significant.
- Section 3.5.3 (Dependency on the ranking score: tf-idf vs. probability of occurrence) showed in an experiment that the top 5 generated terms with the *PVALUE*-method are in most cases also extracted by the *TFIDF*-method in the top ranks. Many terms ranked high using the *TFIDF*-method are not ranked high when the score is calculated using true probabilities (*PVALUE*). On the other side, most of the terms scored high using true probabilities (*PVALUE*) are also ranked high when relying on *TFIDF* derived scores.

In conclusion the stability of the ranking is influenced by the two technical parameters as well as the selection of the global corpus. Tf-idf is a fast approximation to the probability of occurrence, but promotes many terms not extracted using the true probability of occurrence which itself does not promote terms not found using tf-idf. The decision a Part-Of-Speech tagger is not crucial for the ranking, at least for the two systems tested. The use of global corpus has a high influence on the ranking of terms, but surprisingly the differences are not significantly whether a general source based on web sites or a clearly domain specific source like PubMed is used.

Term normalisation

DOG4DAG generated candidate concepts with lexical variations. For this all syntactically or morphologically similar are grouped. This way *DOG4DAG* not only addresses the suggestions made by Nenadić et al. (2004a), who proposed that grouping of lexical variants improves the term recognition result, it also obtains more stable local and global term occurrence counts for improved ranking of candidate terms.

Impact on applications

Overall, the analysis performed in this chapter lead to the conclusion that the proposed term generation method fulfills the specified requirements and is suitable for the integration in ontology engineering tools.

Ontology learning tools

Over the past few years some text-mining approaches and systems for ontology learning such as TerMine, Text2Onto, OntoLT for Protégé, or OntoLearn have been developed.

TerMine based on the C-Value method (Frantzi et al., 2000) retrieves and ranks multi-word phrases. Since 15% of all MeSH terms and synonyms, as well as most gene names are consisting of a single word, *DOG4DAG*'s inclusion of single words as terms is an important extension not present in TerMine. *DOG4DAG* achieves this

by ranking terms according to their relative importance (*tf-idf*). The grouping of all lexical variants and abbreviations leads to better frequency counts and less noise. Text2Onto (Cimiano and Völker, 2005) is an ontology learning framework including a graphical user interface which supports terminology recognition, hypernymic and mereological relationship extraction. The OntoLT Protégé Plug-in (Buitelaar et al., 2004) includes rule based extraction of candidate terms and relations based on linguistic features of provided texts. Both systems build on strong linguistic foundations but require user input prior to the generation of terms or relations, such as the creation of rules in Text2Onto and an annotated corpus of documents for OntoLT (Buitelaar et al., 2004). Our evaluation in Alexopoulou et al. (2008) showed, that the term generation of DOG4DAG performs equally or better than the other state-of-the-art systems Text2Onto and TerMine.

Limitations

There are two major limitations: The ability to compose terms and the ability to generally restrict to a certain aspect in the domain of interest, hence the extraction of a subset of the entire terminology.

Composition of terms Currently, there are many efforts to understand the composition of ontology terms. There are linguistic relationships reflected in term labels like *membrane*, *inner membrane*, *mitochondrial inner membrane* (Ogren et al., 2004) and complex terms like *negative regulation of interleukin 2-biosynthesis* existent in the Gene Ontology (Mungall, 2004). In Aranguren et al. (2008) the authors discussed two design patterns for terms. The term generation method does not support such a composition process. However, filtering as we implemented in the DOG4DAG Ontology Generation Tool in OBO-Edit (Section 7.2) helps to realise the value partition pattern. For example, after a search for "stem cell" one can filter to keep only terms containing "stem cell" obtaining among others the value partition mesenchymal, hematopoietic, and neural.

Subset restriction The acquired terminology is extracted from the text of a specified document corpus. All important words and phrases are extracted. In cases where the ontology engineer wants to concentrate on one specific sort of terminology, e.g. cell lines, organisms, or substances, the proposed method does not provide means for selection. Currently all terminology is extracted, but the applications can provide case-by-case syntactic filtering on the basis of e.g. regular expression filtering to reduce the candidate list to cell types with the expression `blast$|cell$|cyte$`

3.7 Future Work

Automatic variation normalisation As one result of Tsuruoka et al. (2008) it was shown, that fully automatically compiled normalisation rules can extract lexical variants of terms from dictionaries and perform equally well as manually created normalisation rules. The authors are able to show that precision for the look-up of terms only decreases marginally, while recall was improved with each iteration of applying detected rules. This means, that automatically refining lexical variants of terms improves the retrieval performance.

Grouping of terms Better retrieval can be achieved by grouping similar terms beyond lexical or morphological variation. In Section 7.5 (Web-based Term Generation Platform) the terms sharing words are already grouped together and enable a better selection of vocabulary. The performance of such a procedure requires further evaluation.

Detection of siblings On open implementation problem is sibling extraction, the prediction of terms at the same level. This is feasible by extracting list of items from web sites, as e.g. done by Google in its Sets application. Typical list representations on web sites are:

- HTML tags (e.g., , , <DL>, <H1>-<H6> tags).
- Items placed in a table,
- Items separated by commas or semicolons,
- Items separated by tabs.

Definition Extraction

References

Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. In *ISMB (Supplement to Bioinformatics)*, Impact factor 2009: 4.3 (accepted for publication)

A definition extraction method has been designed, implemented and evaluated. The method is capable of extracting and ranking relevant definitional sentences from web search results and web sites. We systematically evaluate each generation step using manually validated benchmarks. Definitions have been extracted for randomly selected 500 GO and 500 MeSH terms. The top 10 ranked definitions per term, in total 10,000 generated definitions, have been manually evaluated to find the best ranked correct or good alternative definitions. For 32% of terms the first extracted definition was correct compared to the original term definition and for 47% of terms it was good but different from the original term definition. For up to 78% of terms good definitions could be retrieved in the top 10 retrieved definitions. No other validated system exists that achieves comparable results. The method has been encapsulated in a web service which enabled the integration in various applications.

The automatic extraction of definitions from natural language text is important for the ontology engineering process. Ontologies are not only a model defining concepts and taxonomic relations – it is further desired that ontologies are a model containing expert knowledge beyond the classification hierarchy. Typically all sorts of relationships (e.g. *is_related_to*, *employed_by*, *visited*, etc.) are used to formulate this knowledge. For a concept itself this expert knowledge is often formulated directly as a definition. In contrast to formal ontologies, biomedical ontologies as the once listed by the OBO Foundry including the Gene Ontology and the Medical Subject Headings do have textual rather than logical definitions. These textual definition describes the concept further by defining the unknown concept through existing concepts. Automatic methods can help to suggest such textual definitions including the reference to literature or web sites containing the definition. We require the generated definitions to follow the pattern *A is a B with property C*. For example the following definition defines the concept “mitosis” through the term “cell cycle process” namely “Mitosis is a cell cycle process comprising the steps by which the nucleus of a eukaryotic cell divides ...”. Further the speciality of “mitosis” in comparison to other

“cell cycle processes” is described by the phrase “*comprising the steps by which the nucleus ... divides*”. Since only organisms with a nucleus in the cell undergo *mitosis* a property restriction defines that the process “*mitosis*” only occurs in the “*eukaryotic cell*”.

Hence the definitions of concepts contain explicit domain knowledge. Partially this knowledge is reflected in the logical structure of the ontology, e.g. the *is_a* relation between “*mitosis*” and “*cell cycle process*”. Other knowledge, like the division of the nucleus in “*two nuclei whose chromosome complement is identical to that of the mother cell*” is often not explicitly formulated in the ontology, but rather remains part of the textual definition.

The support for the creation of textual definitions for term candidates obtained within the ontology learning process has potential to save development time and to create better ontologies with well defined concepts. To achieve this four general requirements have been formulated for a definition generation method.

Requirements

1. The method in general should be **domain independent** to allow the creation of definitions for terms or concepts from diverse knowledge domains.
2. The method should be fast to allow **On-The-Fly interactive generation** of definition candidates.
3. The method should return a **significant number of definitions** to address completeness (recall) for the definitional facts (definiens) obtained for the term to be defined (the definiendum).
4. The method should extract **definitions which contain hyponym relationships** to be incorporated in taxonomy induction task (Chapter 5).

4.1 DOG4DAG Definition Extraction Method

The method was created to be integrated in ontology editors like OBO-Edit and Protégé. Together with the term generation and taxonomy induction the collection of tools and applications will in the following be referred to as Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG).

Method summary

The method aims to generate definitions that will follow the definitional pattern “*A is a B with property C*”, meaning that *A* is defined through the more general term *B* and can be distinguished from other *Bs* by its unique characteristic *C*. For example, *Endocytosis (A) is the process (B) by which cells absorb molecules (such as proteins) from outside the cell by engulfing it with their cell membrane (C)* (from Wikipedia).

For the extraction of definitions four sequential steps need to be performed:

1. Extraction of phrases and sentences containing definitions
2. Annotation of the definiendum (*A*) including resolving ambiguities
3. Annotation of the definiens (*B*) with clear reference to the definiendum
4. Selection and ranking of the definitional statements (*C*) which are answers to the question (for definitional question answering), but are good definitions itself.

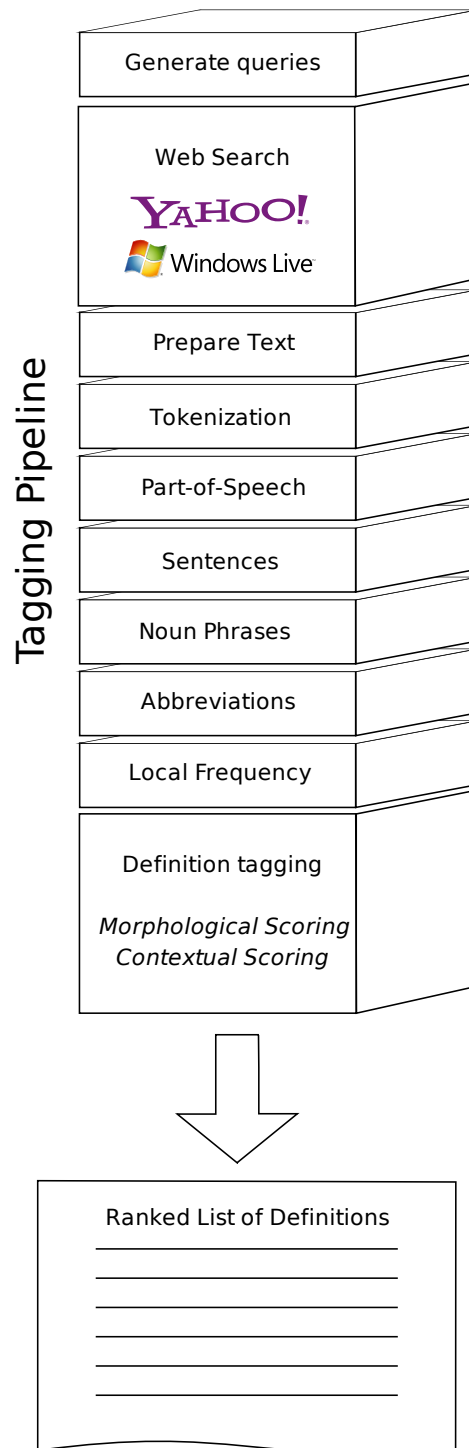


Fig. 4.1. The definition recognition pipeline

Task 1
Phrase Extraction

The task **(1)** finding definition-containing phrases or sentences is designed to use the programmatic interfaces to web search engines which are available to end users. Like Yang et al. (2003); Westerhout and Monachesi (2008); Velardi et al. (2008) we follow the approach to extract definitions from web pages which are a rich source for definitions with high quality especially since dictionaries, lexica, and thesauri are on-line accessible. The term to be defined, denoted as *A*, is used to create queries to retrieve web search results via Yahoo's BOSS and Microsoft's Live Search API. We perform a search for *A* as well as *A* combined with hyponym patterns (Hearst, 1992) of high confidence ("*A is a*", "*A is an*", "*A are*", "*As are*"), or lower confidence ("*such as A*", "*A is*", "*such A like*", "*or other A*", "*and other A*", "*A including*", and "*especially A*"), see Table 4.1. For some queries we restrict the search to sites typically containing definitional statements like *answers.com*, *wikipedia.org*, *reference.com*. Typically 20-40 web searches are performed in parallel to retrieve the definitions for one term. The linked web sites are parsed and the passages around the found keywords are retrieved.

Due to the ever increasing performance of the IT infrastructure also complete mining of large sources became feasible. The estimation is that e.g. for the extraction of all text passages possibly containing definitions from a corpus of approximately 18.000.000 PubMed abstracts a total processing time of 200 CPU hours is needed. PubMed itself is not a good source for definitions. Definitions are typically contained in the full text of an article. The most frequent pattern "*is a*" appears in PubMed abstracts only 75,000 times in the beginning of a sentence not later than the 15th character¹. Assuming an average length for publications is 8 pages and the abstract is 1/16 share of it, the total processing time of all full text articles (if available) could range around 3200 CPU hours, which are only 1.3 days on a 100 CPU cluster.

Task 2
Definiendum
Annotation

The annotation of the definiendum (*A*) **(2)** is similar to the term extraction task as covered in Doms (2009). The author showed recently that the F-measure for the unambiguous identification of an ontological concept in text can reach values of 0.80 and above, by using contextual information in form of positive and negative learning examples. This good result was achieved in a literature mining scenario, where scientific abstracts were selected as learning examples and meta data, such as year, date of publication, or scientific journal was available for the training of the machine learning classifier. Generally the definiendum can be automatically found, but disambiguation is harder then for ontology concepts as learning examples or meta data are rarely available. Very little can be learnt only from the sole term (for definition extraction) or the questions (for definitional question answering). The questions "*Who is ...*" or "*What is ...*" usually only indicate whether a person or a thing has to be defined.

Task 3
Definiens
Annotation

The correct **(3)** annotation of the definiens (*B*) is hard, because the statement made about the definiendum can be a part of the sentence containing the definiendum, can be the whole sentence, or can be spread over several sentences. From the six classes how definitions are defined in web pages (Westerhout and Monachesi, 2008), namely to be (25.5%), verb (30%), punctuation (13.9%), pronoun (13.9%), layout

¹ this estimation of the upper bound of definitional sentences in PubMed assumes that the noun phrase to be defined is no longer than 15 characters

(2.1%), and other patterns (14.5%). According to the categorisation by Westerhout and Monachesi, DOG4DAG searches on 69.4% of the definitions available.

From the high number of candidate definitions suitable definitions are selected and ranked. They are ranked higher according to six criteria regarding the term *A* to be defined and the differentia *C*: First, the definition contains *A* literally; second, the definition starts with *A*; third, *A* is the definition's subject; fourth, *C* starts with an ontology term; fifth, *C* starts with a noun phrase; sixth, the relation *A* is a *B* is found literally. Sentences following the definitional pattern are initialised with score ($P0$) +100 and 0 otherwise. The score is increased when

Task 4
*Selection and
ranking of the
definitional
statements*

- ($T1$) the definiendum is literally contained in the sentence (+30)
- ($T2$) the sentence starts with the definiendum (it is subject in the sentence) (+50)
- ($T3$) the definiendum is subject in the sentence, but not leading phrase (+20)
- ($D1$) a ontology term is leading the differentia (+30)
- ($D2$) a noun phrase is leading the differentia (+20)
- ($P1$) the definitional pattern is *is_a* (+10)

The decision to use explicit scoring is our solution to re-rank definitions on the clients side. Knowledge about the ontology terms is only available on client side i.e. in the ontology editor. $T1$, $T2$, and $T3$ are statements about the definiendum. $D1$ and $D2$ are statements about the differentia. $P0$ and $P1$ are statements about the definitional pattern. $T2$ and $T3$ are mutual to each other. We require $P0 = T1 + T2 + D1$ that in cases where the definitional pattern was not correctly found the definition is extracted but not ranked higher than definitions where a pattern could be found. $T2 > T3$ weights sentences higher, that start with a phrase containing or equal to the term to be defined. $D1 > D2$ achieves that known ontology terms in the beginning of the differentia are weighted higher than other noun phrases. $P1$ promotes all definitions following the "is a" pattern, which is preferably used when definitions are explicitly stated.

To achieve domain independence (**requirement 1**) the method uses Internet search engine results as primary source. The definitions get extracted from the text passages (snippets) typically returned by keyword-based search engines. One limitation regarding the use of snippets obtained from web search engines is, that snippets are usually truncated within sentences. Often only partial definitional sentences can be retrieved. Nevertheless, once a good partial definition is found and annotated in a snippet, the full definition is extracted from the original web site or web document if available. Another limitation is that results are not necessarily reproducible as the index of the web search engine can change. For highly ambiguous terms it can happen that no definitions are retrieved. The reason is that only the top ranked snippets per query (currently the top 1000) are used to find definitions. If the relevant snippets for one sense of the term are ranked lower than the cutoff rank, these snippets are not considered.

Requirement 1
*Domain
Independence*

In web pages, definitions are written in many different ways. Westerhout and Monachesi (2008) investigated this and distinguished between five types how terms in natural language text are defined (see Section 2.3.4 (Generating textual definitions)). In view of a web source for definitions, searchable patterns, like "liver is a" or "endosomes are" seemed to be sufficient for the retrieval of high quality defini-

tional phrases. It was assumed, that somewhere on the web the required information will be given in this form. Especially the possibility to access full text scientific publications via search engines is a big advantage. Even for subscribed journals, the accessibility is limited due to formats and diversity of programmatic interfaces.

This assumption is supported by the evaluation results of Westerhout and Monachesi, where 25% of the definitions were indeed defining the definiendum though a form of “to be”. To increase recall and meet the requirement to retrieve a significant number of definitions (**requirement 3**) and also ensure results for rare topics additional queries for the definiendum are issued to the web search engines.

*Requirement 3
Significant
Number of
Definitions*

In the definition extraction process the definiendum, the definitor and the head nouns of the definiens are annotated within the definition candidates. The candidate sentences are Part-Of-Speech tagged, but no deep linguistic analysis is performed. Even if this additional linguistic information, like dependencies within the sentences, could lead to better filtering and ranking results. Nonetheless the deep linguistic analysis is computationally expensive and counteracts to **requirement 2** for a fast, on-the-fly extraction of definitions. Therefore noun phrases are retrieved based on Part-Of-Speech tagged tokens and a heuristic was developed to decide, whether a definiens truly refers to the definiendum in question. This heuristic simply determines based on punctuation, Part-Of-Speech tags and signal words, if the definiendum is semantically modified by other sentence components or is not the subject of the definition.

*Requirement 2
On-The-Fly
Definition
Extraction*

Regarding the definitor verb, several forms of “to be” and other verbs were annotated to address **requirement 3** to retrieve a significant number of definitions. With respect to the evaluation by Westerhout and Monachesi this would cover already more than 55% of all definitions contained in the corpus. For the annotation of the definition types punctuation, layout and pronoun the quality of the snippets with regard to structure and punctuation were not good enough. The extraction algorithms would be highly prone to errors. Already on proper sentences, the type punctuation definitions were only found with a precision of 0.50 and a recall of 0.36.

*Requirement 3
Significant
Number of
Definitions*

With respect to **requirement 4**, hyponym relationship extraction, web search engines are queried with the definiendum as part of Hearst patterns (Hearst, 1992), e.g. “X, such as Y”, to retrieve more text snippets containing evidence for hyponym relations. High confidence and low confidence patterns have been distinguished. The term to be defined is expanded on the basis of Hearst patterns. The resulting phrases are used to query web search engines.

*Requirement 4
Hyponym
Relations*

Technical implementation

The whole definition extraction pipeline was implemented using the programming language JAVA and was encapsulated in an Axis2 generated web service allowing synchronous and asynchronous requests. This allows the seamless integration into other applications as done for the OBO-Edit Ontology Ontology Generation Plugin (Section 7.2) and the Protégé Term Generation Plugin (Section 7.3). The web service is running on a standard application server with 4 cores and 4GB of main memory together with the services for term generation and ontology look-up. The algorithms are modularised and share common components, like tokenization, noun

<i>High confidence patterns for definitions</i>	<i>Low confidence patterns for definitions</i>
<X> is a	such as <X>
<X> is an	<X> is
<X> are	such <X> like
<X>s are	or other <X>
	and other <X>
	<X> including
	especially <X>

Table 4.1. Examples for Hearst patterns used for definition extraction.

Listing of patterns which indicate for definitional statements with low or high confidence.

phrase extraction, DOM like data structures, abbreviation detection etc. with the term generation module. All system components are managed using MAVEN 2.0, a software project management and comprehension tool. The module was developed using the Open Source IDE Eclipse.

Applications

With the OBO-Edit Ontology Generation Tool (Section 7.2) and the Protégé Ontology Generation Plug-in (Section 7.3) the definition extraction method has been integrated into the most used editors. Both tools have a graphical user interface to display and work with the results provided by the web service.

4.2 Evaluation: Answering TREC2003 definitional questions

Evaluation setup

We evaluated the definition extraction method based on questions and answers of the *TREC2003* task on definitional question answering (Voorhees, 2003). Given a document corpus, this task required participants to find answers for the definitional questions in a defined corpus. In our validation the aim was to show, that searching the web with our definitional patterns and the ranking of retrieved definitional sentences is suitable to suggest valuable definitions. For a definitional question like “Who is Charles Lindberg?” or “What is a golden parachute?” the definitions for the contained *noun phrase* “Charles Lindberg” and “golden parachute” have been generated. Like in the original task the required answer is a the list of sentences, in this case a list of definitions.

For 50 questions (Table 10.5) the generated definitions have been manually compared with the answers given by the assessors board. Sentences containing facts marked in the original benchmark as “vital” or “ok” have been labeled as accepted answers for this evaluation. Two example questions are listed in Tables 4.3 and 4.4 together with the expected answers and the first three generated definitions. For the first example question “What are fractals?” the first answer refers to a software called “Fractals” and second answer to a company with the name “fractal”. The third answer was judged correct, because “self similarity” and “all scales” from the accepted answer are mentioned. The answers for the second example “What is the

<i>Range</i>	<i>correct</i>	<i>correct %</i>
correct top 1	20	40%
correct within top 5	37	74%
correct within top 10	45	90%
no correct	5	10%

Table 4.2. Evaluation results for answering TREC2003 definitional questions.

For 40% of the 50 questions as correct answer could be found as top extraction result. For 90% of the question correct answers could be found without consulting the text corpus provided in the original task.

vagus nerve?” were all judged correct. The “vagus nerve” has been correctly identified as a “nerve” in the “body” connecting “brain”, “heart”, ..., “stomach”, etc., so connects “internal organs”.

Results

For 20 questions out of 50 (40%) the top candidate definition was a correct definition. In 74% (37/50) of the cases a correct definition was found in the top 5 and in 90% (45/50) a correct definition could be found in the top 10 terms. For only 5 questions the method failed to find correct definitions. The reason for missing the definition was manifold:

- **Correct definition hidden in popular content:** *Akbar the Great* – The correct answers have been contained in the result list, but famous sites as his tomb dominate the web search results. The practice of extending the search term relying on Part-Of-Speech fails for “Akbar the Great” and also leads to many statements which are too complex to be identified as good definitions, e.g. “Akbar, arguably the great Mughal emperor is a paragon of perfection!”. An additional difficulty is that Akbar is not always referred to as “Akbar the Great”, e.g. “Akbar is considered as the great Mughal emperor who put ...”.
- **Disambiguation:** *Anthony Blunt* – With the whole web as source the name “Anthony Blunt” could not be correctly disambiguated.
- **No definition contained in web search results:** *Abraham in the Old Testament* – The retrieved texts did not contain a definition of “Abraham in the Old Testament”. “Abraham” as name itself is widely used and could not be disambiguated.
- **No text has been found:** *Ph in biology* or *the medical condition shingles*

These results are in line with the top competition results with 0.21 precision of the best system (Liu et al., 2003) (see Table 4.2). See Chapter 10 (Appendix) Table 10.10 for all questions and the position of correct definitions in the result set. The full set of answers and manual curations are listed in Table 10.5.

Qid 1957: What are fractals?

Accepted answers

fractals – is a pattern that is irregular but self-similar at all size scales

Generated answers

- 1 Fractals is a Java-based product providing rules-based and intelligent card fraud detection for all types of payment card, for both card issuers and card acquirers
www.alaric.com/public/products/fractals
 - 2 Fractal is a leading provider of advanced analytics that helps companies leverage data driven insights in making better decisions.
<http://www.fractalanalytics.com/>
 - ✓ 3 fractal is an object or quantity that displays **self-similarity**, in a somewhat technical sense, **on all scales**.
<http://mathworld.wolfram.com/Fractal.html>
-

Table 4.3. Example of generated answers for questions from TREC2003 definitional question answering task. For the questions Qid 1957: “What are fractals?” the third has been judged as correct.

Qid 2008: What is the vagus nerve?

Accepted answers

vagus nerve – extends from the brain stem to most of the body’s internal organs

vagus nerve – it relays orders from the brain to regulate things like heart rate, while keeping the brain

vagus nerve – informed about what’s going on in the organs, such as whether the stomach is full

Generated answers

- ✓ 1 The vagus nerve is a major **parasympathetic nerve that meanders and branches through the body, from the anterior brain through the esophagus, trachea, heart, diaphragm, stomach, and ...**
<http://www.healingtouchyoga.com/pranayama.html>
 - ✓ 2 The vagus nerve is an important nerve of the parasympathetic nervous system and **innervates several organs in the neck, thorax and abdomen.**
http://bss.evi.utwente.nl/people/students_master/thiele_kobus.doc/index.html
 - ✓ 3 The vagus nerve is **a nerve that carries messages to and from the brain - connected to internal organs such as the heart and stomach.**
http://www.ing.md/s_vagal_stim.htm
-

Table 4.4. Example of generated answers for questions from TREC2003 definitional question answering task. For question Qid 2008: “What is the vagus nerve?” the first three answers have been judged as correct.

4.3 Evaluation: Generation of GO and MeSH definitions

The comparison to *TREC2003* (Section 4.2) is encouraging and allows to compare the method to the state-of-the-art, but does not cover the life sciences. For a specific evaluation of biomedical ontologies, we compared the generated to manually created definitions.

Evaluation setup

To assess how well generated definitions can approximate manually created definitions, we randomly selected 500 GO and 500 MeSH terms (Tables 10.3 and 10.4) manually verified whether generated definitions matched the GO/MeSH definition or in another case gave useful information. In our opinion 500 terms per ontology are sufficient for this evaluation and constitute an amount that is still feasible to be manually evaluated. For the 500 GO and 500 MeSH terms 10 definitions were generated and manually labelled as either *correct* if they match the GO/MeSH definition or *good* if they were at least sensible and relevant. All generated definitions are listed in Tables 10.6 and 10.7. A definition was judged as *correct* if it followed the original GO/MeSH definition with structure “*A is a B with property C*” by at least agreement in *B* followed by a reasonable good *C*, or alternatively agreement in *C*, given a reasonable good *B*, typically a more general or specific term than the original *B* (see examples in Table 4.5). If generated definitions matched the GO/Mesh definition exactly they were excluded since the likely source was the original definition. This happened 5 times out of 10,000 definitions. Since GO terms rarely appear literally in text, see e.g. (Ogren et al., 2004), definitions for GO terms have been evaluated excluding common pre- and postfixes. E.g. for “*myosin binding*” we generated for “*myosin*” the definition “*Myosin is a protein possessing multiple functions integral to muscle contraction, force generation, muscle development, and production of high-quality processed meats.*”, which we compared to the original GO definition. We excluded the pre- and postfixes “*activation*”, “*activity*”, “*binding*”, “*regulation of*”, “*localization*”, “*development*”, “*transport*”, “*catabolic process*”, “*metabolic process*”, and “*biosynthetic process*”. This applied to 307 of the 500 GO terms. The quality for definition extraction is measured in terms of *precision*, *recall*, and *f-measure*.

Results

On the whole, nearly all GO and MeSH terms have definitions with an average of 24 words (GO) and 30 words (MeSH) and contain 2.4 ontology terms (GO) and 5.7 (MeSH) (see Table 4.6). Hence the GO and MeSH provided a good benchmark for definition generation. All 10,000 generated definitions (10 for each of the 500 GO and 500 MeSH terms) were manually verified whether they matched the GO/MeSH definition (*correct*) or were proper definitions of acceptable quality (*good*). A number of example definitions and whether they were considered as *correct* or only *good* are provided in Table 4.5. The complete list of all generated definitions is given in Tables 10.6 and 10.7. The top definition was in over 40% *good*, meaning that it was a proper definition containing useful information about the term. The results increased to 55% (GO) and 78% (MeSH) for the top 10 definitions (see Table 4.7). Over half of these 78% were actually *correct* definitions showing that the automated definitions are by and large of acceptable quality for interactive ontology generation.

Original		Generated	
Gene Ontology			
integrin biosynthetic process (GO:0045112)	The chemical reactions and pathways resulting in the formation of integrins, a large family of transmembrane proteins that act as receptors for cell-adhesion molecules.	4th: integrin is a heterodimer transmembrane protein that plays a critical role in cellular adhesion and migration during the inflammation and immune response. [eng.umd.edu]	correct
anion channel activity (GO:0005253)	Catalysis of the energy-independent passage of anions across a lipid bilayer down a concentration gradient.	1st: Anion channel is an integral membrane protein or more typically an assembly of several proteins. [cogsci.uni-osnabrueck.de]	good
benzoate metabolic process (GO:0018874)	The chemical reactions and pathways involving benzoate, the anion of benzoic acid, a fungistatic compound widely used as a food preservative; [...]	1st: Benzoate is a common carbon source in nature that is funnelled directly to the widely distributed benzoyl-coenzyme A (benzoyl-CoA) central pathway. [mic.sgmjournals.com]	good
cerebral cortex development (GO:0021987)	The progression of the cerebral cortex over time from its initial formation until its mature state. The cerebral cortex is the outer layered region of the telencephalon.	1st: cerebral cortex is a layer of nerve cells forming a convoluted outer shell over the brain, [...] in which much of the thinking or higher intellectual activity of the brain takes place. [www.hermes-press.com]	good
Medical Subject Headings			
Flucytosine (D005437)	A fluorinated cytosine analog that is used as an anti-fungal agent.	1st: Flucytosine is a fluorine analog of cytosine [...], leading to inhibition of thymidylate synthetase and disruption of DNA synthesis. [medicine.medscape.com]	correct
Cystoscopy (D003558)	Endoscopic examination, therapy or surgery of the urinary bladder.	3rd: cystoscopy is an examination of the bladder [...] using a flexible, miniature telescope [...] [www.nuffielhealth.com]	correct
Xanthomonas campestris (D016959)	A species of gram-negative, aerobic bacteria that is pathogenic for plants.	1st: Xanthomonas campestris is a Gram-negative plant-pathogenic bacterium [...] [mic.sgmjournals.org]	correct
Trypanosoma brucei gambiense (D014347)	A hemoflagellate subspecies of parasitic protozoa that causes Gambian or West African sleeping sickness in humans. The vector host is usually the tsetse fly (Glossina).	1st: Trypanosoma brucei gambiense is a blood borne, flagellated protozoan which is transmitted to humans and animals via the tsetse fly (Glossina spp.). [etd.lib.ttu.edu]	correct

Table 4.5. Original and the best generated definition for 4 GO and 4 MeSH terms.

Definition are manually labelled as either *correct* if they match the GO/MeSH definition or *good* if they contain useful information. For each generated definition the rank of retrieval (1st, 2nd, 3rd, or 4th) is shown.

	All GO	All MeSH
Total	28814	29348
Terms with definition	99.1%	96.0%
Words in definition	24.3% ($\pm 15.3\%$)	30.2% ($\pm 19.3\%$)
Terms in definition	2.4% ($\pm 2.3\%$)	5.7 ($\pm 4.1\%$)
≥ 1 term in definition	88.0%	97.2
≥ 1 ancestor in definition	54.1%	56.2
≥ 1 parent in definition	15.8%	36.6

Table 4.6. Proportion of terms from in MeSH and GO containing parent terms, ancestor terms or other existing terms in their definitions. Nearly all of GO an MeSH terms are defined. 54.1 – 56.2% of terms are defined via an ancestor, 15.8 – 36.6% via a parent term.

	500 GO		500 MeSH	
	<i>correct</i>	<i>good</i>	<i>correct</i>	<i>good</i>
Top 1	21.9%	41.2%	32.0%	47.0%
Within top 2	24.6%	47.8%	41.6%	60.2%
Within top 5	27.8%	54.6%	49.8%	72.6%
Within top 10	27.8%	54.6%	53.6%	78.2%

Table 4.7. Evaluation of generated definitions for 500 GO and 500 MeSH terms.

For 22 – 38% of terms the top ranked definition captured aspects of the true definition, in 41 – 47% it was a good definition, but not similar to the original one. Within the top 10 ranked definitions a good definition was found for 55 – 78% of terms.

4.4 Summary and Discussion

A definition extraction method has been specified and developed. The domain independent method has been evaluated using a benchmark of the 50 questions from the definitional question answering task held at *Text REtrieval Conference (TREC) 2003*. The evaluation shows that definitional statements can be extracted for general terminology.

The definition extraction method has been evaluated large scale for terms in the life sciences. Definitions have been generated for 1,000 randomly selected terms from two frequently used resources, the Gene Ontology (GO) and the Medical Subject Headings (MeSH). For each of the both resources 500 terms have been subject to automatically find definitions. To gain confidence in the method the high number of 10,000 definitions, the top 10 for each term, have been evaluated by hand and compared to the existing definition of the terms in GO and MeSH.

The evaluation on three independent benchmarks shows that a good definition can be found within the top 10 predictions in 88% (TREC, definitional question answering), 54% (GO) and 78% (MeSH) of cases. In terms of recall an evaluation is not possible as this would require manual annotation of all candidate definition available on-line. The results show that definitions for life science terminology can be extracted from web content. A fully automatic extraction of definitions as formulated in the hypothesis associated with research question 1 (Section 9.1) is not yet feasible. The top ranked definition is currently a good definition for 41% (GO) to 47% (MeSH) of terms.

The evaluation based on TREC shows similar results as yielded in the original competition. Let's assume that for each term at least one definition is available and that only the first definition would have been submitted. Under this assumption the F-measure for extracting this answer is 0.4. Compared with the original competition results of 0.27 - 0.31 (Table 2.12) this is in the same range, with the difference that the extraction is based on web content, an unrestricted source of information, not on a limited corpus of topic-specific sentences as used for TREC. The F-measure in the original evaluation is lower as for some question more than one definitional fact was required to be found. This leads to a reduction of recall.

Regarding the extraction of definitions for medical terms e.g. Velardi et al. (2008) achieves 0.76 precision and 0.36 recall in a small scale example of 17 terms (see Table 2.11). A second evaluation by Velardi et al. lead to 0.85 precision and a coverage of 96%, meaning that for nearly all terms one good positively evaluated definition could be found.

For the task of finding just one good definition our evaluation lead to a lower precision of 0.47. As we randomly selected terms this lower precision was expected. Firstly, not all terms in MeSH can be expected to have been defined somewhere in text (e.g. "Immunoglobulin Km Allotypes", "Fluids and Secretions"). Parts of these terms can certainly be defined automatically. Secondly, not all labels of terms in MeSH correspond to the used terminology typically stored as synonyms. In this evaluation known synonyms have not been considered.

Hence, the important difference in the mode of evaluation between Velardi et al. (2008) and DOG4DAG is, that Velardi et al. evaluates the generation of definitions for terms previously generated from text by their system, while DOG4DAG has been evaluated on randomly chosen terms from existing controlled vocabularies which have not been extracted from text. It also does not become clear under which criteria a definition has been judged as correct in the evaluation by Velardi et al.. It has been noted by Velardi et al. that these good results might vary significantly depending on the knowledge domain.

The high number of 78% of terms with a good definition in the top 10 generated definitions and the approx. 55% of definitions which contain an ancestor term in the original definitions (Table 4.6) suggest that definitions in the life sciences are a rich source for taxonomic relations. This will be further investigated in Section 5.1 (DOG4DAG Taxonomy Generation Method).

Limitations

We identified three limitation for the proposed definition extraction method. First, snippets obtained from web search engines are usually truncated within sentences. Often only partial definitional sentences can be retrieved. Nevertheless, once a good partial definition is found and annotated in a snippet, the full definition is extracted from the original web site or web document if available. Second, results are not necessarily reproducibility as the index of the web search engine can change. Caching could be a solution, but has not been integrated. Third, for highly ambiguous terms it can happen that no definitions are retrieved. The reason is that only the top ranked snippets per query (currently the top 1000) are used to find definitions. If the relevant snippets for one sense of the term are ranked lower than rank 1000, these

snippets are not considered. All definition will then be for the sense of the term contained on web sites which have been ranked high.

Glossary generation

For many technical terms the glossary entries in the end of the thesis have been generated or extended using the definition extraction method which has been defined, developed and evaluated as part of this thesis Chapter 4 (Definition Extraction).

The generated glossary entries are labeled with the symbol



4.5 Future Work

The work on definition extraction was initiated with the intent to find existing definitions to support the process of defining ontology concepts. Currently several definitions are being retrieved which often state distinct facts about a term. Future methods need to go further to automatically find one definition. Therefore such methods will need to include:

1. Disambiguation of the term defined in the context of a definition. The context can be provided by close ontology terms or as described by Saggion and Gaizauskas (2004) using co-occurring terms
2. Consideration of synonyms of terms to find definitions
3. Decomposition of facts from definitional sentences
4. Composition of full definitions compiling several facts
5. Composition of definitions containing cross sentence border information by using co-reference information (Saggion and Gaizauskas, 2004)
6. Systematic revision of relevance measures for definitional patterns (Ravichandran and Hovy, 2002)
7. Integration of existing glossaries and structural filters to find definitions in web documents as used by Velardi et al. (2008)

Taxonomy Generation

References

Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. In *ISMB (Supplement to Bioinformatics)*, Impact factor 2009: 4.3 (accepted for publication)

Henschel, A., Wei Lee Woon, Wächter, T., Stuart Madnik (2009). Comparison of generality based algorithm variants for automatic taxonomy generation. In *Proceedings of 6th International Conference on Innovations in Information Technology*, December 15-17, AlAin, United Arab Emirates.

Wächter T., Tan, H., Wobst, A., Lambrix, P., and Schroeder, M. (2006). A corpus-driven approach for the design, evolution, and alignment of ontologies. In *Proceedings of Winter Simulation Conference (Computational Systems Biology)*, Monterey, CA, USA (3rd–6th December 2006), pages 1595–1602, Invited contribution.

A method to generated taxonomic relations on the basis of generated definitions has been designed, implemented, and evaluated for 1,000 randomly selected ontology terms from GO and MeSH. For 54% of terms in MeSH and 38% of terms in GO, correct relations to ancestors could be predicted. Additional two experiments have been performed to test the suitability of pattern-based and statistical methods for finding parent child relations. Based on a data set of approximately 200,000,000 term occurrences in PubMed abstracts it has been experimentally shown that the taxonomic relations of the Medical Subject Headings (MeSH) can be reconstructed with a $F\text{-measure}_{0.5}$ of only 0.13 using co-occurrence of terms in very large corpora, whereas for distinct branches of MeSH like Tissues, Blood, Sense Organs, Virus Diseases could be predicted with an $F\text{-Measure}_{0.5}$ between 0.42 and 0.60.

Taxonomies as hierarchies of classes are the basis for classification. They are widely used to structure information in the life sciences and form the taxonomic backbone of ontologies. Taxonomy generation aims to hierarchically arrange concepts (or classes) in a automatic manner. In literature two types of methods have been described, namely **lexico-syntactic methods** and **statistical methods** (Cimiano, 2006b). The task has also been referred to as the creation of a noun hypernym hierarchy (Caraballo, 1999) or concept hierarchies (Cimiano et al., 2005).

Taxonomies provide the information needed to generalise or specialise, let it be on the basis of data – by finding subsets or super sets according to the relation supporting the taxonomy – or the level of concepts – to find more specific or general

terms and categories. From a modelling perspective a representation of knowledge in taxonomies enables the specification of axioms, rules, relations, and queries in a defined and efficient way by making redundant the enumeration of all concrete classes.

Ontology creation has been motivated in Section 2.1 as difficult, labour intensive and target for automation. This certainly also applies to the creation of taxonomies (or concept hierarchies). It will be investigated to what extend biomedical taxonomies can be created using pattern-based methods on the basis of previously extracted textual definitions and to what extend statistical methods based on co-occurrence information are able to predict the taxonomic relations defined in biomedical taxonomies.

Motivated by the integration of the method in interactive ontology engineering applications a number of requirements have been defined for the task taxonomy generation.

Requirements

1. **Domain independence:** The method in general should be **domain independent** to allow the creation of subsumption relationships between terms or concepts from diverse knowledge domains.
2. **Performance:** The method should be fast to allow **On-The-Fly interactive generation** of subsumption relationships.
3. **Precision:** The automatic extraction of subsumption relationships needs to be performed with **high precision**. For ontology learning all relationships need to be manually validated by an expert. A higher proportion of correctly predicted relationships and hence a better taxonomic structure makes this validation of relationships less difficult.
4. **Coverage:** For the method recall is less important than precision, although high recall is desirable.
5. **Transparency:** The method should be **transparent**. It has been observed in manual annotation projects that human judges often do not agree on annotations or validation results – the inter-annotator agreement is low. To make it possible to re-think manual annotation, the methods should allow the collection of evidence for a relation. Ideally references to trusted sources, let it be scientific literature or established databases, should be collected for each predicted relationship.

5.1 DOG4DAG Taxonomy Generation Method

Adhering to the specified requirements a taxonomy generation method has been defined, developed, and evaluated. Taxonomy generation, i.e. finding parent-child relationships, is an easy problem if one has a definition of the form “*A is a B with property C*”, where *B* is the parent of *A*. As Table 5.1 shows, nearly all definitions in GO and MeSH mention at least one term. But it also shows that only 16% (GO) and 37% (MeSH) contain the parent in the definition. However, it increases to over 50% when *B* is not necessarily the parent but an ancestor of the defined term. Interestingly, some of the sub-ontologies, namely *organism* (MeSH), *anatomy* (MeSH), *geography* (MeSH) and *cellular component* (GO) provide much better results with values of

	All GO	All MeSH
Total	28814	29348
Terms with definition	99.1%	96.0%
Words in definition	24.3% ($\pm 15.3\%$)	30.2% ($\pm 19.3\%$)
Terms in definition	2.4% ($\pm 2.3\%$)	5.7 ($\pm 4.1\%$)
≥ 1 term in definition	88.0%	97.2
≥ 1 ancestor in definition	54.1%	56.2
≥ 1 parent in definition	15.8%	36.6

Table 5.1. Proportion of terms from in MeSH and GO containing parent terms, ancestor terms or other existing terms in their definitions. Nearly all of GO an MeSH terms are defined. 54.1 – 56.2% of terms are defined via an ancestor, 15.8 – 36.6% via a parent term (*same as Table 4.6*)

over 70%. For a detailed break down see Tables 10.8 and 10.9. As a consequence, our method uses generated definitions as source for predicting parents.

Method summary

Given definitions of the form “*A is a B with property C*” we extract existing terms similar to *B* as candidate parents in a parent-child relationship. Terms are regarded as similar if they show a Hamming distance of less than 20% of the length of the shorter term label or synonym. The Hamming distance between two strings denotes the number of position the two strings differ from each other. We align strings from the beginning and include the length of overlapping tails in the distance. All ontology terms are ranked starting with the identical term, known parents from other ontologies, predicted parents from confirmed definitions, predicted parents from generated definitions, and finally terms syntactically similar to the term to define.

Applications

This method has been implemented as part of the ontology generation tool introduced in Section 7.2 (OBO-Edit Ontology Generation Tool).

5.2 Evaluation: Taxonomy generation based on generated definitions

Evaluation setup

For 500 random GO and 500 random MeSH terms (Tables 10.3 and 10.4) definitions have been generated (Tables 10.6 and 10.7). In the evaluation we investigate to what extent known parent or ancestor terms are literally contained in the top 10 generated definitions for each term. Syntactic variation is not considered for locating known terms in the definitions.

Results

Based on the 10,000 definitions generated for 1,000 terms in Section 4.3 (Evaluation: Generation of GO and MeSH definitions), we tested how many definitions contain the parent or an ancestor of the term the definition was generated for (Table 5.2). For 13% (GO) and the 26% (MeSH) the top 10 generated definitions contained the direct

parent term, thus DOG4DAG will predict it. For 38% (GO) and 54% (MeSH) the top 10 generated definitions contained an ancestor term, thus DOG4DAG will predict some correct, but indirect, ancestor relationship. For the vast majority of GO and MeSH terms already the top ranked generated definition contains the parent/ancestor. The numbers for ancestors within the top ten generated definitions correspond nicely to the manual curations in Table 4.7, with a *correct* definition in the top ten for 28% of the GO and 54% of the MeSH terms.

	500 GO		500 MeSH	
	<i>parent</i>	<i>ancestor</i>	<i>parent</i>	<i>ancestor</i>
Contained in top 1	12.2%	32.4%	20.2%	37.0%
Contained in top 10	13.4%	38.0%	26.0%	54.4%

Table 5.2. Evaluation of taxonomic information contained in generated definitions for 500 GO and 500 MeSH terms. For 26% of the 500 randomly selected MeSH terms the parent and for 54% some ancestor could be found in the top 10 generated definitions.

5.3 Pattern-based relation extraction – Superstring prediction

In (Ogren et al., 2004) the compositional structure of GO terms was analysed. The authors found that many GO terms contain each other and many GO terms are derived from each other. For example, the term *membrane* [GO:0016020] has *inner membrane* [GO:0019866] as a direct sub-concept. This knowledge can be used to automatically generate new candidate terms following the observed patterns. We analysed, whether these super-string relations observed in GO, can be verified in the text.

Evaluation setup

By analysing the GO we identified 3,129 out of 20,223 terms, where the term is a super-string of its children. Further for 1,189 of these terms (6% of all GO terms), the term and its children were found in PubMed abstracts (see also Table 5.3). Based on at most 5,000 texts containing the parent terms we identify the words which precede the actual term and rank them by their frequency of occurrence. This lead to a list of newly identified candidate terms to be possibly included in the ontology. In the following we show the generated candidate child terms for the example GO terms “Death”, “vacuole”, and “GTPase activator activity”. Valid prefixes and prefixes matching existing child terms contained in the GO are indicated in the column GO. Many of the predicted terms are children of the known parent term.

Terms in GO	20223
Terms found in abstracts	14905
Terms having children containing themselves	3129
... parent found in text	2692
... parent and one child found in text	2239
... parent and one child found in text; parent substring of child	1189
... parent and all children found in text	1781
Terms having children	7451
... parent found in abstracts	5964
... parent and one child found in abstracts	5185
... parent and all children found in abstracts	3757

Table 5.3. Statistic on Gene Ontology terms appearance in PubMed abstracts with and without their known child terms. 74% of all 20223 GO terms (as of Dezember 2005) can be found in PubMed abstracts; 29% of terms can be found and have children; 26% of terms can be found in text and at least one child term can be found; 19% of terms can be found and all children can be found in text. Hence 26% of terms it is theoretically possible to infer an parent child relationship on the basis of PubMed, which is the upper bound for the method described in Section 5.4 (Co-occurrence analysis – Algorithm by Heymann et. al). For 6% of terms the parent is substring of the child and both are contained in text which is the upper bound for the method described in Section 5.3 (Pattern-based relation extraction – Superstring prediction).

Example: GO:0005096 ‘GTPase activator activity’

GTPases are molecular switches. A GTPase activator is an enzyme that catalyzes the hydrolysis of GTP. GTPase activator activity has the children ‘ARF’, ‘Rab’, ‘Rac’, ‘Ral’, ‘Ran’, ‘Rap’, ‘Ras’, ‘Rho’ and ‘Sar GTPase activator activity’. Five of the children can be automatically found.

pos.	candidate	count	in GO
1	ras	133	child
2	rho	106	child
3	small	100	similar term
4	intrinsic	88	
5	gap	37	synonym
6	p21ras	34	
7	family	29	
8	arf	23	child
9	triphosphatase	19	similar term
10	rac	17	child
11	p21	16	
12	rab	12	child
..

Example: GO:0016265 'Death'

This term has the children 'aging', 'tissue death' and 'cell death'. Out of these three terms the superstring prediction method is only capable to find 'tissue death' and 'cell death'. While 'cell death' is found first, 'tissue death' is not found within the first 50 predicted terms. Nevertheless by carefully investigating the result list one will find, that many terms are from the medical domain rather than molecular biology. Terms like 'cardiac death', 'neuronal death', 'infant death', 'fetal death', 'brain death' and 'neonatal death' make perfectly sense for a medical ontology. Predicted prefixing words like 'sudden', 'early' and 'late' can easily be filtered using knowledge about their frequency of occurrence in the English language.

pos.	candidate	count	in GO
1	cell	60678	child
2	sudden	11521	
3	cardiac	7179	suggested child
4	neuronal	5326	suggested child
5	infant	3925	
6	fetal	3636	suggested child
7	brain	3468	suggested child
8	early	2658	
9	late	2079	
10	neonatal	2038	suggested child
..

Example: GO:0005773 'vacuole'

A vacuole is defined as a closed structure, found only in eukaryotic cells, that is completely surrounded by unit membrane and contains liquid material. The term has the children 'autophagic', 'contractile', 'lytic', 'parasitophorous' and 'storage vacuole'. All are found in the first 50 predicted terms.

pos.	candidate	count	in GO
1	autophagic	1219	child
2	cytoplasmic	1048	suggested child
3	parasitophorous	933	child
4	large	684	
5	food	496	
6	contractile	387	child
7	phagocytic	383	suggested child
8	rimmed	383	
9	lipid	378	suggested child
10	intracellular	303	suggested child
11	intracytoplasmic	295	suggested child
12	digestive	265	descendant
13	endocytic	260	suggested child
14	small	247	
15	membrane-bound	240	suggested child
..
20	storage	175	child
..
44	lytic	36	child
..

Results: Superstring prediction

For the experiment only those terms were considered, where at least one child and its parent term were contained in text, and where the parent term was literally contained in the child term. The analysis was performed separately for two cases, where either

- (a) the child term ends with the parent term (1062 of 1189 cases), or
- (b) the child term starts with the parent term (127 of 1189 cases).

Per parent term a maximal number of 5,000 PubMed abstracts have been analysed and term occurrences have been counted. The terms preceding or subsuming the parent term have been ranked by frequency of occurrence. The hypothesis saying that parent terms are contained in child terms as proper sub-string has been shown to hold for many biomedical terms in more than 15% of the cases (3129 of 20223 GO terms). Ogren et al. (2004) reported that $A \subset B$ given A is parent of B in 25.5% of the cases (4,197 of 16,451 GO terms). Although the composition of terms is a pattern in the Gene Ontology, in the experiments it was not investigated to what extent string inclusion can be found in other domains. The method is domain independent as no domain specific information is required. The method is simple and fast and can be easily integrated in interactive learning tools. The OBO-Edit Ontology Generation Plug-in as well as the Protégé Plug-in provide a regular expression filter functionality which allows finding candidate according to experiment (a) with a query “<child> <parent>\$” and candidates according to experiment (b) with “<parent> <child>\$”.

Requirement 1
Domain
independence

Requirement 2
Performance

(a) Child term ends with parent term			(b) Child term starts with parent term		
	top 5	top 10		top 5	top 10
children found (of 1062)	276	334	children found (of 127)	35	43
recall	26.0%	31.5%	recall	27.6%	33.9%
precision	6.9%	4.1%	precision	0.9%	0.5%
maximal precision	26.4%	13.2%	maximal precision	3.2%	1.6%

Table 5.4. Precision and recall observed for the top 5 and top 10 ranked potential child terms for the cases where the child terms (a) ends with and (b) starts with the parent term.

The results of the analysis in terms of precision and recall are shown in (Table 5.4). The simple experiment shows on average very low precision of less than 10% for experiment (a) and even lower than 1% for experiment (b). The overall precision is expected to be higher when including noun phrase chunking for filtering to allow only valid noun phrases as child terms. Although it can be expected that the true precision will be higher as valid candidate terms which are not part of the test resource (in this case the Gene Ontology) are regarded as false predictions. With respect to requirement 4 Table 5.4 shows a recall of 26% for experiment (a) and 27.6% for experiment (b). The method is transparent in a way, that all terms extracted from text can reference the texts they have been extracted from. Nonetheless there is no

Requirement 3
Precision

Requirement 4
Coverage

Requirement 5
Transparency

explicit evidence, and hence no transparency for the assignment of subsumption relationships.

5.4 Co-occurrence analysis – Algorithm by Heymann et. al.

Co-occurrence data has been used to predict taxonomic relations but performance was generally lower than 0.5 in terms of precision and recall (Sanderson and Croft, 1999; Witschel, 2005; Ryu and Choi, 2006). It will be analysed whether larger data sets available for life science literature can be used for taxonomy generation. The data set was obtained from GoPubMed¹ and contains all manual and automatically created assignments of MeSH terms to the approximately 18 million scientific abstracts listed in PubMed.

The used algorithm relies on the hypothesis that term y is a child of term x , if term y occurs in a subset of the document set in which term x occurs. Such a parent-child relationship is assigned only if the conditional probability of $P(x|y)$ and $P(y|x)$ satisfy conditions like $P(x|y) \geq \text{threshold}$ and $P(y|x) < 1$. The Heymann and Garcia-Molina algorithm (Table 5.5) uses a simplification where in a first step all terms are ordered according to their centrality in the similarity graph. In a second step terms are assigned to a graph either as top element or as child of a previously assigned term under the condition, that the similarity between terms x and y exceeds a defined threshold. One drawback of the method is in the assignment of nodes to at most one parent term. Biomedical ontologies as our examples Gene Ontology and MeSH often make use multiple parent assignments. The method has been implemented using two different measures for centrality. BETWEENNESS centrality and mean distance to all other vertices in the similarity graph (CLOSENESS centrality).

Let $G : (V, E)$ be a similarity graph, where V is a set of contained vertices and E the set of edges between the vertices. Let cosineSim be a binary function (the cosine similarity) which defines the similarity between vertices in V .

BETWEENNESS
CENTRALITY

Definition 5.1 (BETWEENNESS centrality). The BETWEENNESS centrality $C_{\text{BETWEENNESS}}$ of a vertex $v_i \in V$ is defined by the number of shortest paths through v_i between any two other vertices's in V .

$$C_{\text{BETWEENNESS}}(v_i) = \sum_{s,v,t \in V; s \neq v \neq t; s \neq t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

with σ being the number of shortest paths from s to t , and $\sigma_{st}(v_i)$ the number of shortest paths from s to t that pass through a vertex v_i . The complexity is $O(n^3)$.

CLOSENESS
CENTRALITY

Definition 5.2 (CLOSENESS centrality). The closeness centrality $C_{\text{CLOSENESS}}$ of a vertex $v_i \in V$ is defined as the sum of the average distance of v_i to all other vertices in the similarity graph. The complexity is $O(n^2)$.

$$C_{\text{CLOSENESS}}(v_i) = \frac{1}{\sum_{v_j \in V \setminus \{v_i\}} 1 - d(v_i, v_j)}$$

¹ <http://www.gopubmed.org>

² according to http://en.wikipedia.org/wiki/Centrality#Betweenness_centrality

Require: $L_{generality}$ is a list of tags t_i, \dots, t_j in descending order of their centrality in the similarity graph.

Require: Several functions are assumed: $s(t_i, t_j)$ computes the similarity (using cosine similarity, for example) between t_i and t_j . $getVertices(G)$ returns all vertices in the given graph, G .

Require: $taxThreshold$ is a parameter for the threshold at which a tag becomes a child of a related parent rather than of the root.

```

1:  $G_{taxonomy} \leftarrow \langle \emptyset, root \rangle$ 
2: for  $i = 1 \dots |L_{generality}|$  do
3:    $t_i \leftarrow L_{generality}[i]$ 
4:    $maxCandidateVal \leftarrow 0$ 
5:   for all  $t_j \in getVertices(G_{taxonomy})$  do
6:     if  $s(t_i, t_j) > maxCandidateVal$  then
7:        $maxCandidateVal \leftarrow s(t_i, t_j)$ 
8:        $maxCandidate \leftarrow t_j$ 
9:     end if
10:  end for
11:  if  $maxCandidateVal > taxThreshold$  then
12:     $G_{taxonomy} \leftarrow G_{taxonomy} \cup \langle maxCandidate, t_i \rangle$ 
13:  else
14:     $G_{taxonomy} \leftarrow G_{taxonomy} \cup \langle maxCandidate, root \rangle$ 
15:  end if
16: end for

```

Table 5.5. An extensible greedy algorithm for hierarchical taxonomy generation from social tagging systems using graph centrality in a similarity graph of tags (transcript from Heymann and Garcia-Molina (2006)).

Evaluation setup

In order to be able to construct subsumption hierarchies for terms from MeSH and Gene Ontology all abstracts containing the term have been extracted from PubMed. The Heymann and Garcia-Molina algorithm is used to generate these hierarchies. The experiment has been performed for the entire MeSH taxonomy as well as for a number of sub branches.

Tested generated sub branches of MeSH and GO

- “cellular_component” from GO,
- “metabolic_process” from GO,
- “enzyme regulator activity” from GO,
- “Tissues” from MeSH,
- “Blood” from MeSH,
- “Cardiovascular System” from MeSH,
- “Sense Organs” from MeSH,
- “Virus Disease” from MeSH, and
- all of Medical Subject Headings.

It was tested to what extent the subsumption relationships defined in MeSH 2007 and GO 2008 could be re-constructed based on the calculate co-occurrence statistics. The experiments have been performed with varying parameters

Parameters

- Maximal number of documents retrieved per term:
1k, 10k, 100k, 1M, 2M, 5M, 10M, and 18M documents
- Threshold for the assignment of a node to the graph or root:
0.5, 0.4, 0.3, 0.2, 0.1, $5 * 10^{-2}$, 10^{-2} , 10^{-3} , 10^{-4} , and 10^{-5}
- Calculation of the centrality for each node using the similarity measures *BETWEENESS* and *average sum of distances*

Analysis steps

1. Retrieval of annotations from www.GoPubMed.org
2. Creation of document vectors with a maximal length of 18M PubMed entries
3. Creation of the similarity network
4. Calculation of the centrality

Evaluation steps

1. Creation of statistics containing the information listed in Table 5.7
2. Creation of charts *number of documents* vs. *F – measure_{0.5}*
3. Creation of graphs in XGMML format
4. Annotation of the XGMML graph with information on the correctness of edges
5. Visualization in Cytoscape

Results: Algorithm by Heymann et. al.

The result for the generation of the whole MeSH graph are listed in Table 5.7. The maximal observed precision for the prediction of relationships between all MeSH vertexes was 0.27. Weakening the criteria for correct predictions and hence considering that indirect relations can be regarded as correct relations ($A..B$), and assuming that direct relations with inverse direction (BA) can be repaired manually, than the maximal achieved precision was 0.34. Recall on the other side is very low between 0.02 and 0.03. The lower the threshold *thresh* is chosen, the more relations are predicted, while precision decreases and recall increases. The maximal F-measure $f_{0.5}$ was observed with 0.14 and 0.21 for the weaker criteria (for F-measure see definition 2.4).

Higher results have been obtained for selected sub branches from MeSH and Gene Ontology. An example of the prediction of the sub branch Blood from MeSH is shown in Figure 5.7.

In the following for each sub branch the best observed F-measure ($f_{0.5}$) will be listed. It will be shown which threshold lead to this F-measure and how many documents occurrences have been used per term to calculate pairwise similarity between terms

Results for tested sub branches from MeSH and GO Compared to the results by Sanderson and Croft (1999), who found 48% correct relations, and Snow et al. (2006), who achieved a precision of 0.58 the results for the different sub branches of

Label	Ontology	$f_{0.5}$	N	thresh	Reference
"cellular_component"	GO	0.38	10M	0.10	Figure 5.1(a)
"metabolic_process"	GO	0.38	18M	0.10	Figure 5.1(b)
"enzyme_regulator_activity"	GO	0.37	2M	0.05	Figure 5.2(a)
"Cardiovascular System"	MeSH	0.42	2M	0.20	Figure 5.2(b)
"Tissues"	MeSH	0.52	1M	0.05	Figure 5.3(b)
"Blood"	MeSH	0.60	1M	0.05	Figure 5.3(a)
"Sense Organs"	MeSH	0.42	1M	0.05	Figure 5.5(a)
"Virus Disease"	MeSH	0.48	2M	0.05	Figure 5.5(b)

Table 5.6. Performance of the co-occurrence based generation of taxonomic relations for selected sub branches from GO and MeSH. The best example for the sub branch "Blood" from MeSH reaches an F-measure $F_{0.5}$ of 0.6.

GO and MeSH lead to similar results, even though a much bigger corpus was used. Taxonomy generation based on co-occurrence alone is therefore not the method of choice for high quality taxonomic relationship prediction.

Results for different centrality measures In Henschel et al. (2009) we analysed the performance and the complexity of two variants of the Heymann algorithm using betweenness and closeness centrality combined with a systematic threshold evaluation on a set of four branches. Betweenness and closeness can be calculated using weighted and unweighted graphs, both variants have been investigated. Figures 5.4 and 5.6 show the results for the four example MeSH branches Blood, Tissues, Sense Organs, and virus diseases. Unweighted betweenness centrality generally performs best but often only marginally better than the faster unweighted closeness centrality. Exception is the network for Blood where the weighted centrality measures performed better. A good choice for the threshold τ_s is $0 < \tau_s \leq 0.1$. The best value varies between the experiments. In further experiments in Henschel et al. (2009) experimented with different variants of the algorithm. Re-ranking the centrality after insertion of a node improves precision in 46% of the cases, decreases precision in 23% of the cases and achieves equal precision in 30% of the cases. Filtering nodes with entropy > 0.7 improves the precision for some branches significantly: "Blood" (from 0.60 to 0.81), "Carbohydrates" (from 0.38 to 0.43), and "Fungi" (from 0.31 to 0.39).

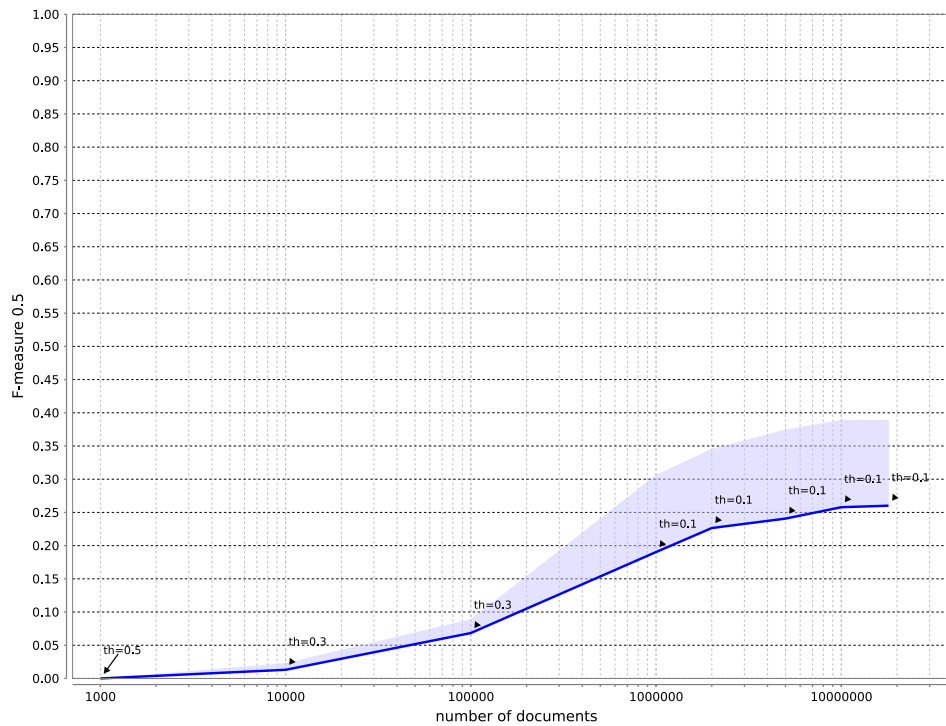
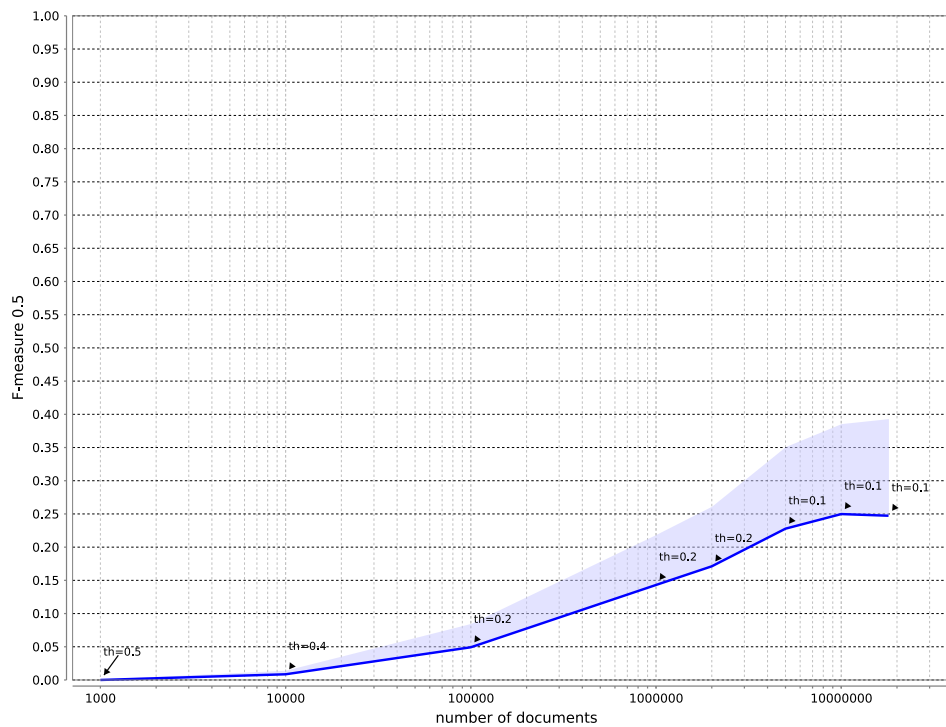
(a) GO branch *cellular_component*(b) GO branch *metabolic_process*

Fig. 5.1. Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for GO branches “cellular component” and “metabolic process”. The area above the chart shows the expected precision if the direction of relations and relations with other terms in between are treated as correct predictions.

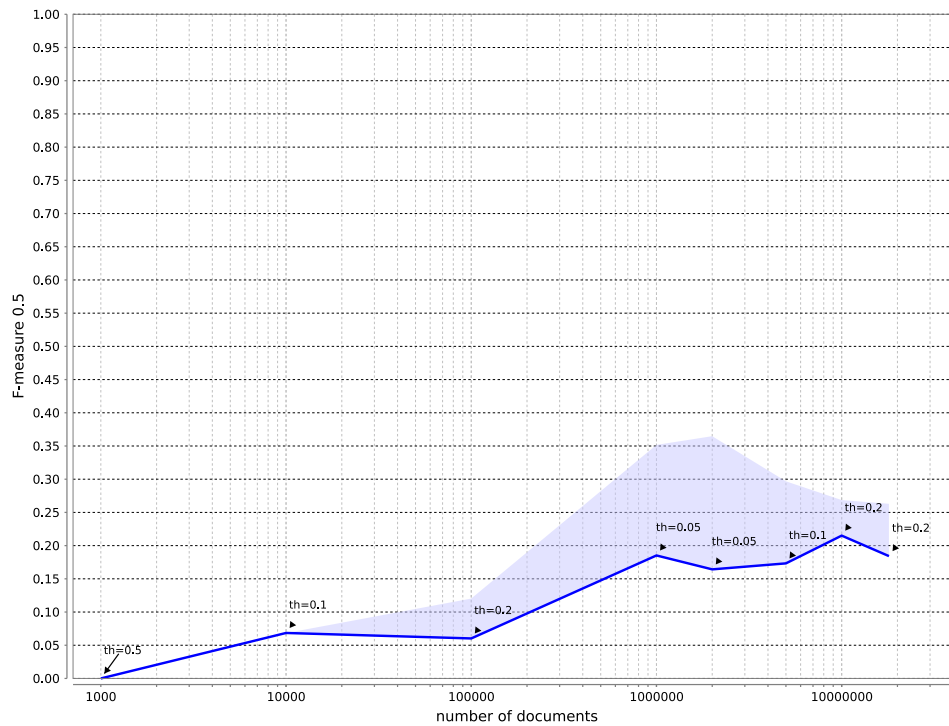
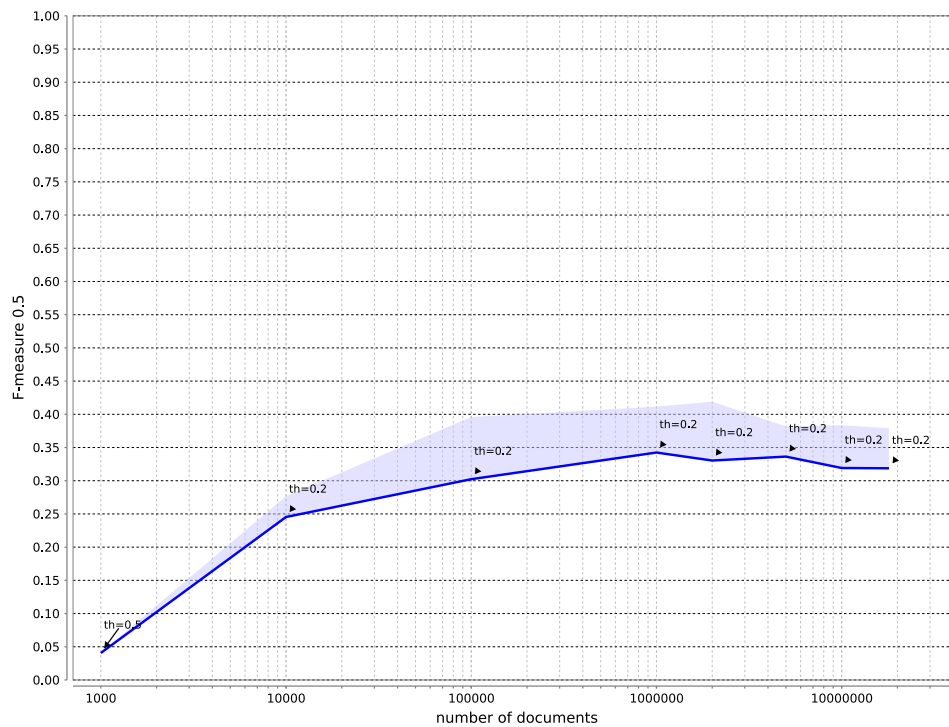
(a) GO branch *enzyme regulator activity*(b) MeSH branch *Cardiovascular System*

Fig. 5.2. Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Examples for GO branches “enzyme regulator activity” and “Cardiovascular System”. The area above the chart shows the expected precision if the direction of relations and relations with other terms in between are treated as correct predictions.

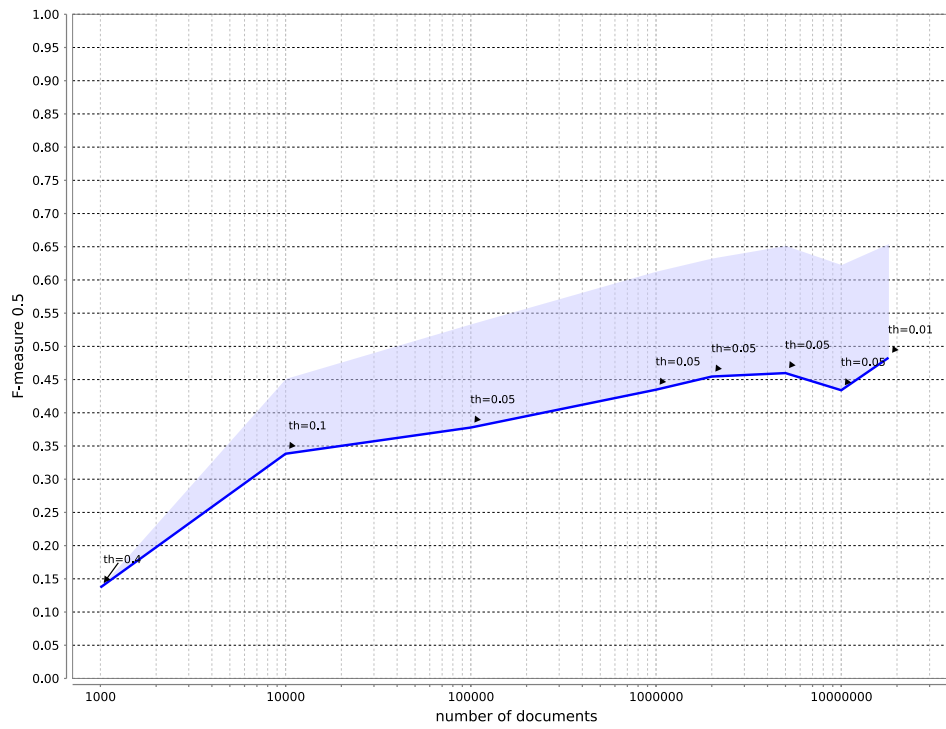
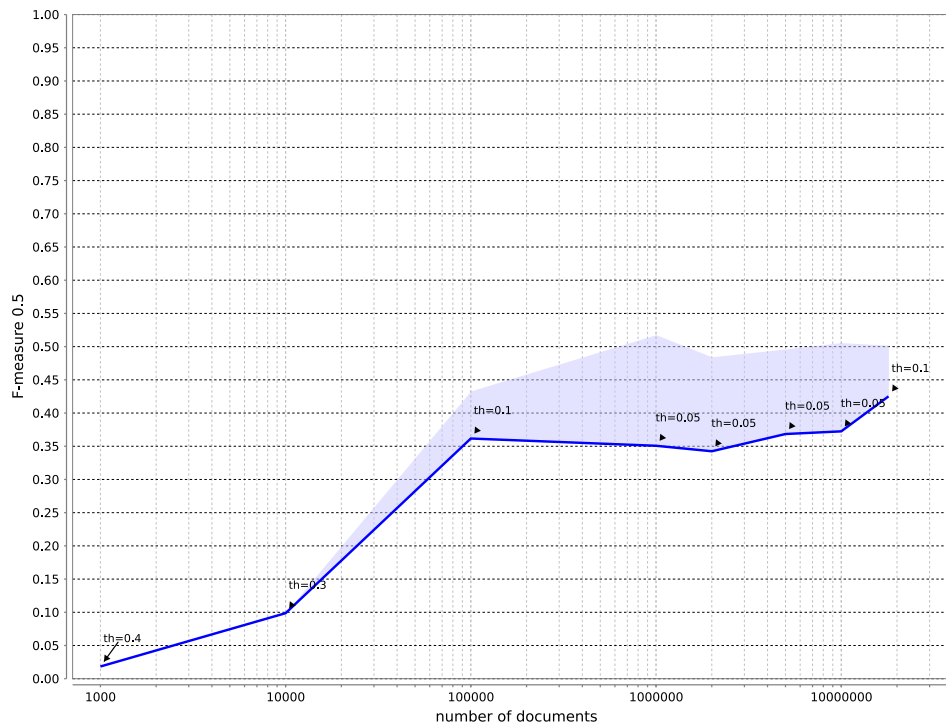
(a) MeSH branch *Blood*(b) MeSH branch *Tissues*

Fig. 5.3. Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for the MeSH branches “Blood” and “Tissues”. The area above the chart shows the expected precision if the direction of relations and relations with other terms in between are treated as correct predictions.

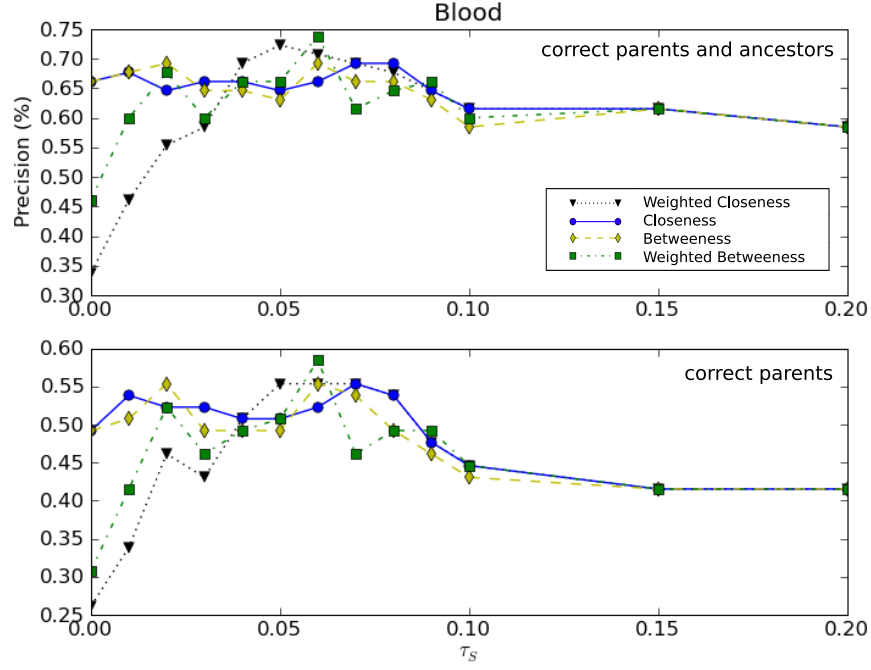
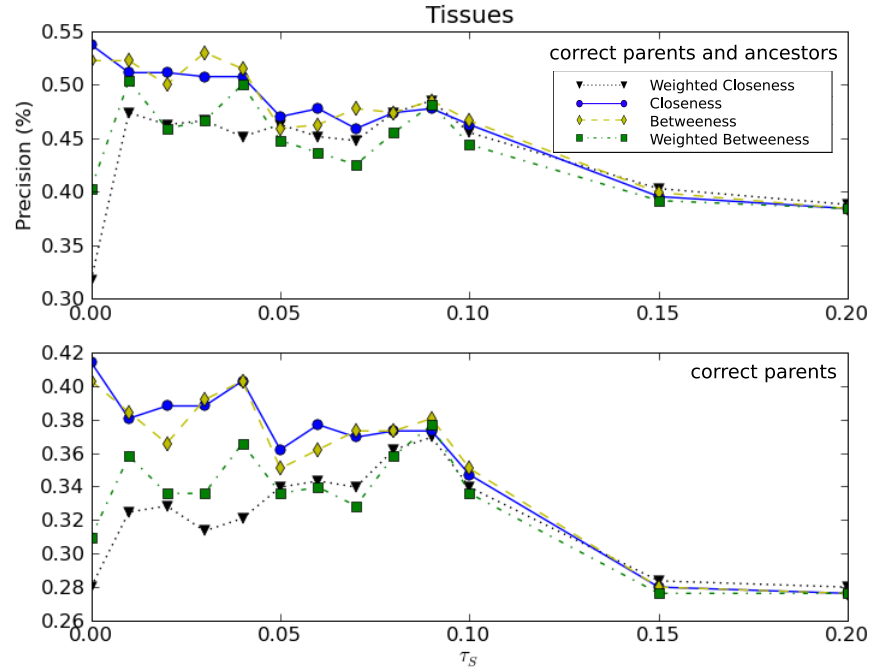
(a) MeSH branch *Blood*(b) MeSH branch *Tissues*

Fig. 5.4. Precision curves for centrality variants MeSH branches “Blood” and “Tissues” A threshold of $0 < \tau_S < 0.1$ lead to the best precision in both branches with > 0.7 (ancestors correct) and > 0.55 (parents correct) for branch Blood. For branch Tissues the results are lower. The centrality measure weighted closeness and weighed betweenness perform best for single thresholds for the Blood network, otherwise the unweighted variants perform better (Henschel et al., 2009).

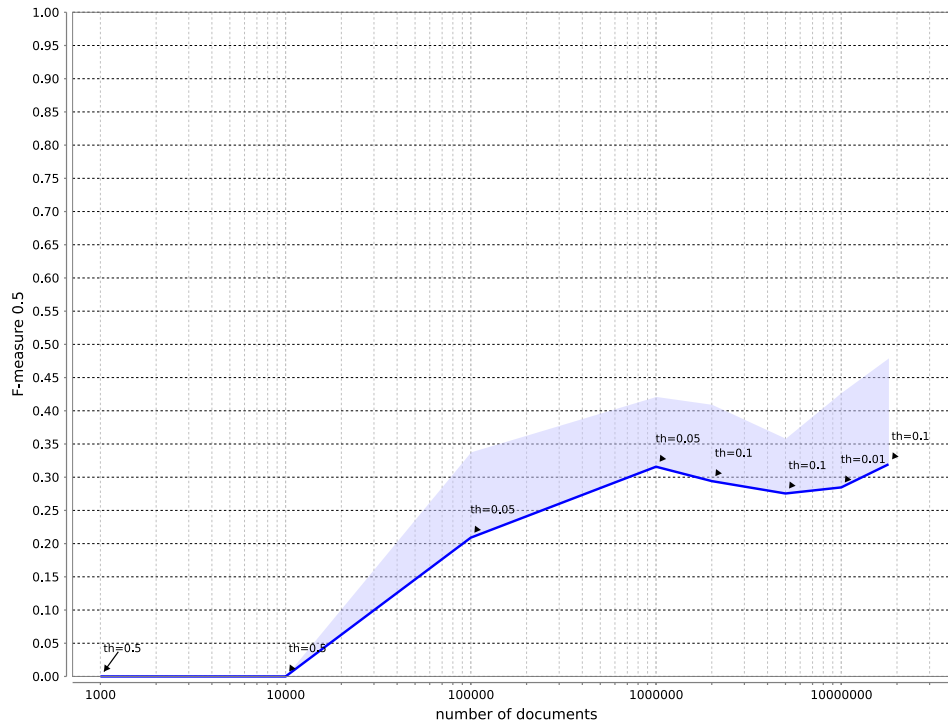
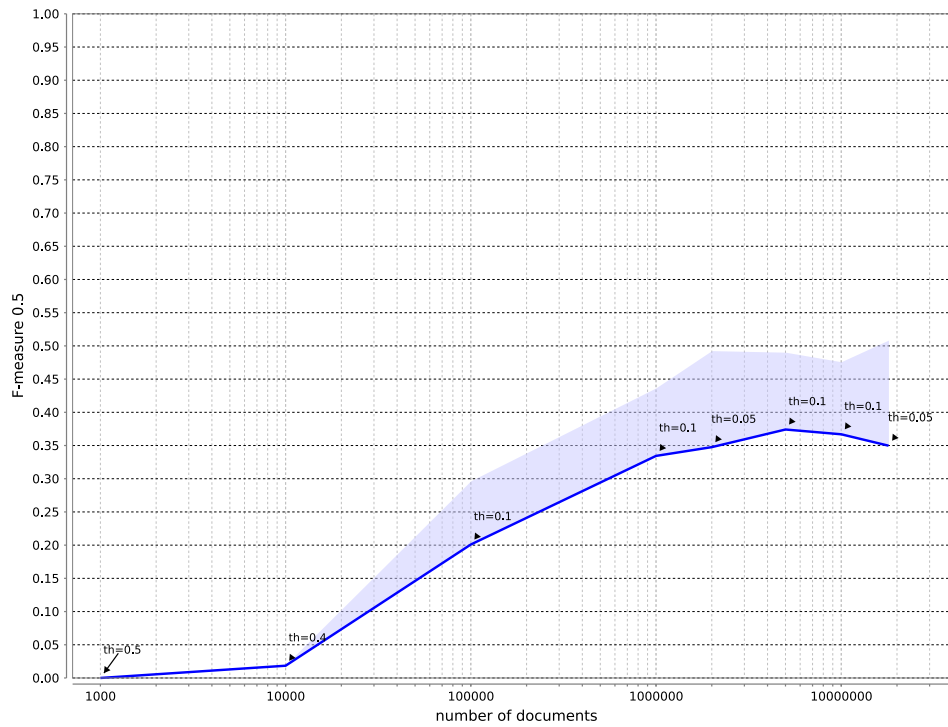
(a) MeSH branch *Sense Organs*(b) MeSH branch *Virus Disease*

Fig. 5.5. Performance of co-occurrence based taxonomy generation against a threshold and the maximal number of documents considered per node. Example for the MeSH branches “Sense Organs” and “Virus Disease”. The area above the chart shows the expected precision if the direction of relations and relations with other terms in between are treated as correct predictions.

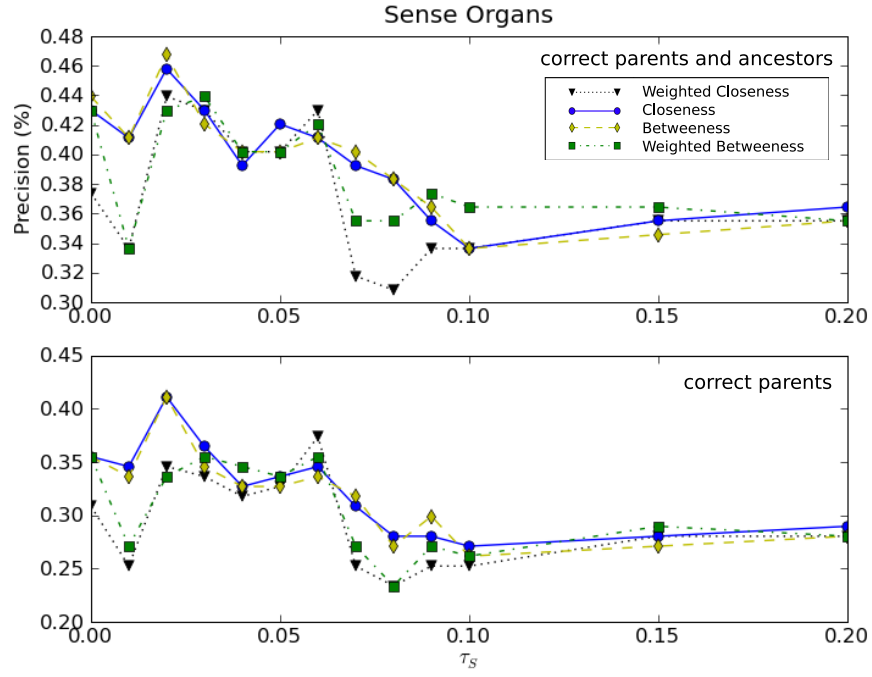
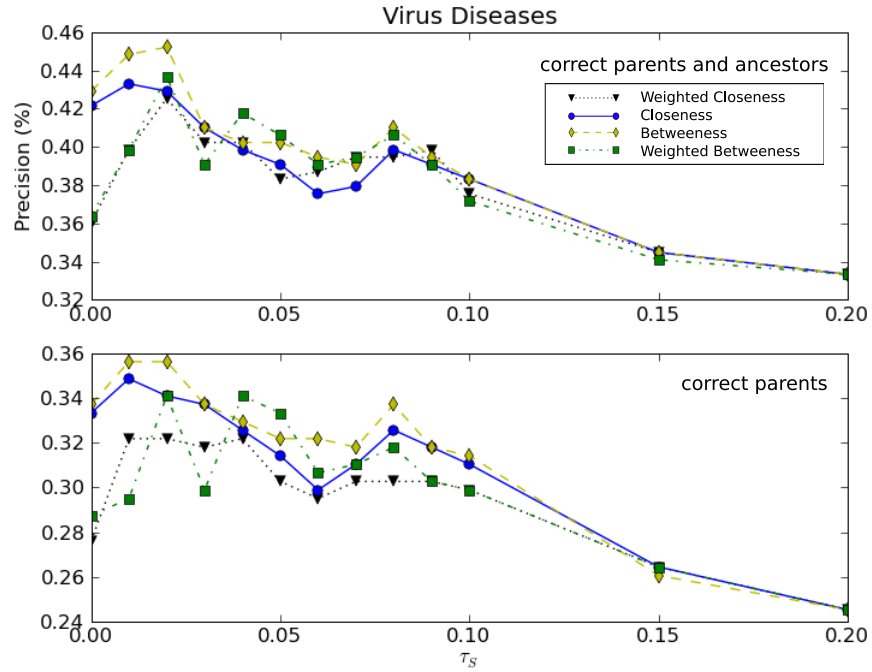
(a) MeSH branch *Sense Organs*(b) MeSH branch *Virus Disease*

Fig. 5.6. Precision curves for centrality variants for the MeSH branches “Sense Organs” and “Virus Disease” A threshold of $0 \leq \tau_S \leq 0.1$ lead to the best precision in both branches with 0.35-0.40 (parents correct) and 0.45-0.46 (ancestors correct). The centrality measure closeness and betweenness perform best for both branches. (Henschel et al., 2009)

Number of correct predicted relations contained in MeSH

	<i>thresh</i>	<i>roots</i>	<i>relations</i>	<i>AB</i>	<i>BA</i>	<i>A..B</i>	<i>B..A</i>
MeSH 0.5		18690	1570	431	73	34	1
MeSH 0.4		17956	2610	679	89	63	7
MeSH 0.3		16625	4455	1046	142	124	19
MeSH 0.2		13458	8607	1661	269	282	40
MeSH 0.1		5706	18170	2641	520	612	93
MeSH 0.05		717	23899	3014	620	754	139
MeSH 0.01		1	24697	3047	635	767	150
MeSH 0.001		1	24697	3047	635	767	150
MeSH 0.0001		1	24697	3047	635	767	150
MeSH 0.00001		1	24697	3047	635	767	150

Precision & Recall

	<i>thresh</i>	<i>precision</i>	<i>recall</i>	<i>F – measure</i>	<i>precision*</i>	<i>recall*</i>	<i>F – measure*</i>
MeSH 0.5		27.45	1.85	3.47	34.33	2.32	4.34
MeSH 0.4		26.02	2.92	5.25	32.11	3.6	6.48
MeSH 0.3		23.48	4.5	7.55	29.88	5.72	9.6
MeSH 0.2		19.3	7.14	10.42	26.16	9.68	14.13
MeSH 0.1		14.53	11.35	12.75	21.28	16.61	18.66
MeSH 0.05		12.61	12.95	12.78	18.94	19.45	19.19
MeSH 0.01		12.34	13.09	12.7	18.62	19.76	19.18
MeSH 0.001		12.34	13.09	12.7	18.62	19.76	19.18
MeSH 0.0001		12.34	13.09	12.7	18.62	19.76	19.18
MeSH 0.00001		12.34	13.09	12.7	18.62	19.76	19.18

Legend:**General statistic counts**

<i>label</i>	label of the term
<i>N</i>	number of document containing the term
<i>thresh</i>	algorithm specific threshold for minimal similarity
<i>roots</i>	number of terms to become direct child of ROOT
<i>relations</i>	number of relations found
<i>AB</i>	correct prediction of relationship
<i>BA</i>	prediction of relationship with inverse direction
<i>A .. B</i>	prediction of indirect relationship
<i>B .. A</i>	prediction of indirect relationship with inverse direction

Quality measures

<i>precision</i>	percentage of correct predicted relations within all predicted relations
<i>recall</i>	percentage of predicted relation from all existing relations
<i>F-measure</i>	F-measure for the predicted relations
<i>precision*</i>	percentage of “possibly correct” predicted relations within all predicted relations
<i>recall*</i>	percentage of “possibly correct” predicted relations from all existing relations
<i>F-measure*</i>	F-measure for the predicted relations regarding “possibly correct” relations as correct

Table 5.7. Results for the reconstruction of sub-class relationships existing between 23,270 nodes in MeSH 2007. The results show maximal precision of (a) 27.45% with recall at 1.85%, and (b) 34.33% with recall at 2.32% when considering direct relations which have been predicted as inverse *BA* and indirect *A..B* relations as correct. The maximal F-measure for case (a) 12.78% and (b) 19.19%.

5.5 Summary and Discussion

In this chapter three methods for taxonomy generation have been evaluated. An own method Section 5.1 (DOG4DAG Taxonomy Generation Method) has been developed and evaluated large scale. Additional experiments with pattern-based and co-occurrence based approaches were performed to test the suitability of the methods for applications in the life sciences.

Taxonomy from definitions (Section 5.2) A method has been defined which extracts relations from definitions obtained by web searches. Definitions are a rich and reliable source for taxonomic relations, as a term is usually defined via more general terms. With respect to **Research Question 1** *To what extent can ontology construction be automated?* (Section 9.1) it has been shown on a representative subset of 1,000 randomly chosen biomedical terms in Chapter 4 (Definition Extraction) that definitions can be suggested for 78% of them. On the basis of definitions the taxonomic structure of ontologies can be predicted.

Evaluation of taxonomy generation results In order to evaluate our approach in the life science domain we need to assess whether automatically learned parent-child relationships are correct. In the area of ontology learning Maedche and Staab (2002) discussed the evaluation of learned ontologies by comparing the lexical and taxonomic overlap between two ontologies. The taxonomic overlap is calculated as the similarity between concepts based the set of all super and sub concepts (ancestors and descendants), by only regarding concepts present in both ontologies and excluding the compared concept itself. This elegant and general measure can be used for fully generated ontologies. In our evaluation we quantitatively evaluate the correctness of the predictions as done by Hearst (1992) and Caraballo (1999), but instead of the number of correctly predicted relationships we determine for how many terms correct relationships can be found. Our evaluation of generated definitions for a representative subset of 500 randomly selected MeSH terms and their existing parent-child relations in Section 5.2 (Evaluation: Taxonomy generation based on generated definitions) revealed, that valid parent and ancestor terms can be predicted for up to 54% of terms in MeSH. The method retrieves a ranked list of potential parent terms. To interpret the results in terms of precision and recall an assumption on the number of definition has to be made. Considering only the predictions based on the top ranked definition, a correct parent term can be found for 20% of the terms an ancestor can be found for 37% of the cases. Assuming that only one parent term exists per term, this is equivalent to a precision and recall of 0.2, or 0.37 for indirect correct relationships. In an application which considers the top ten ranked definitions as source for taxonomic relationships a parent term can be found for 26% of terms, an ancestor for 54%. This means that the recall is 0.54, again under the assumption that only one parent exists per term. Precision has not been measured in the experiment. For a meaningful precision calculation all generated definitions need to be manually evaluated with respect to the described relationship. For the 500 randomly selected MeSH terms this would be 5,000 definitions. In Section 4.3 (Evaluation: Generation of GO and MeSH definitions) we manually curated these definitions to find the first correct definition in the retrieved list but did not qualify the contained relationship.

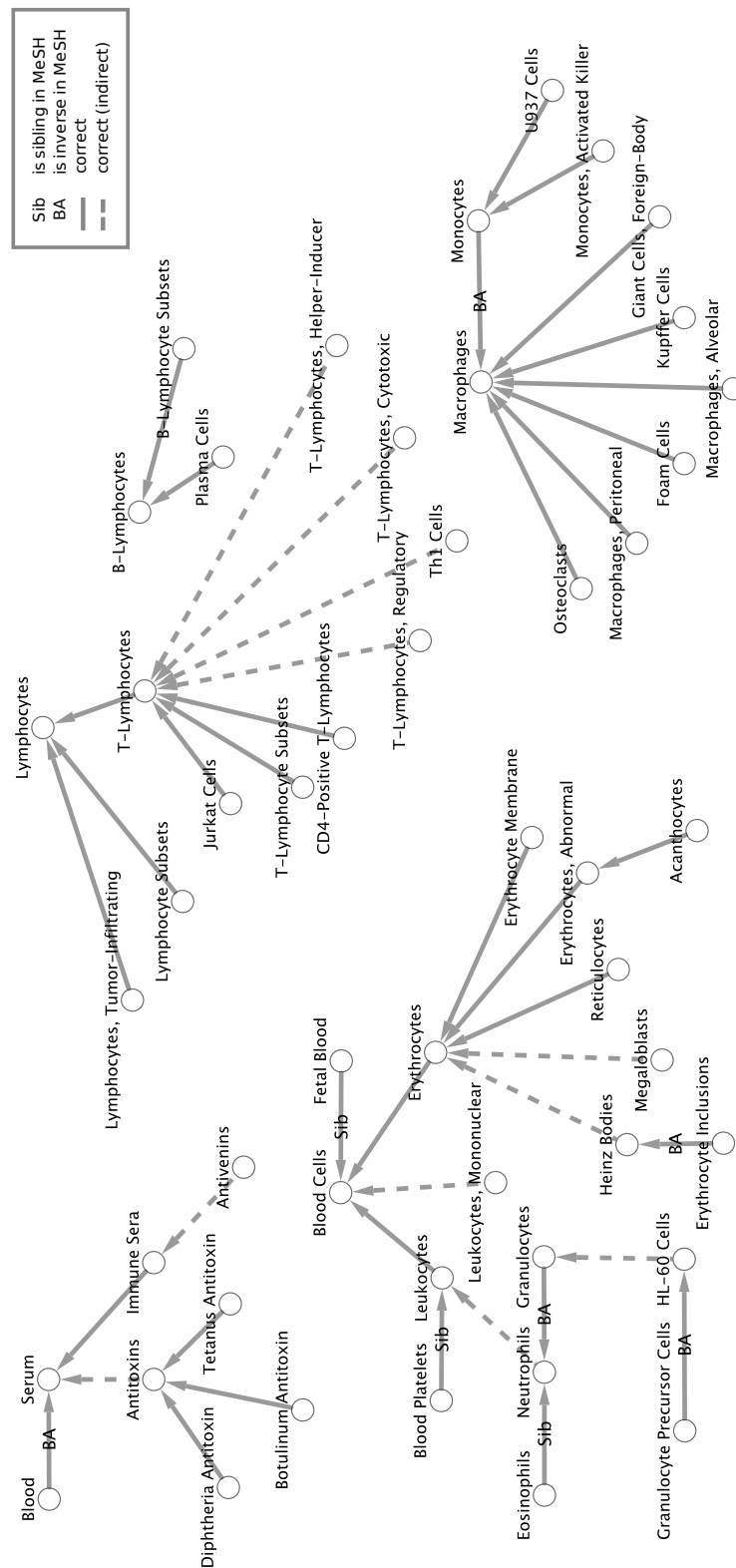


Fig. 5.7. Generated taxonomy graph sub branch "Blood" in MeSH using co-occurrence based taxonomy generation. The graph was created with the method described in Section 5.4 (Results: Algorithm by Heymann et. al) using a maximum of 10,000,000 documents per node and a threshold of 0.01.

Semi-automatic taxonomy generation An important advantage of definition-based taxonomy generation is the availability of reference information and the definitional statements for all suggested relationships. This makes the prediction transparent and revisable. Explicit evidence can be given for each predicted relation. The revision of the relationship is easier, because the definitions supporting the statement are known and can be presented to a human curator who can curate the relation in the context it has been formulated.

Additional experiments

Pattern-based string inclusion and co-occurrences have shown the ability to retrieve parent child relations, but are vulnerable to over generation of candidate children. The evaluation lead to less than 20% precision for the prediction of taxonomic relationships from text. The method using string inclusion analyses n-gram frequencies without using linguistic information. With all word pairs potentially being terms, the method used was too simple to reliably capture the relationships in the GO. Co-occurrence builds on statistical information but does not provide direct causal evidence for predicted relationships.

Taxonomy from string inclusion (Section 5.3) The evaluation of the purely pattern-based *Superstring analysis* revealed that the information of string inclusion does not achieve high recall, which was also reported for pattern-based methods by Hearst (1992); Nenadić et al. (2004a) and others. But, simple string inclusion fails to predict parent child relationships with high precision. We achieved in this initial experiment a precision of 0.07 when considering prefixes and 0.01 when considering suffixes. As candidate predictions are ranked in a list only the top five predictions have been considered to calculate precision and recall. The maximal precision achieved in a single experiment was 0.26 for prefix suggestion and 0.03 for suffix suggestion. A summary of the results is provided with Table 5.4.

Taxonomy from co-occurrence (Section 5.4) To investigate the performance of *co-occurrence* as information for the prediction of taxonomic relations we implemented a version of the algorithm proposed in Heymann and Garcia-Molina (2006) and tried to recover taxonomic relations between the more than 20,000 terms defined in the Medical Subject Headings (MeSH). In over 18,000,000 scientific abstracts listed in the literature database PubMed we annotated occurrences of terms and obtained the document-wise co-occurrence for all $20,000 \times 20,000$ pairs of terms from MeSH based on approx. 200,000,000 term occurrences in PubMed. From this we tried to re-construct the subsumption relationships. For the prediction of all relationships existing in MeSH a F-measure of only 0.13 was reached. Precision and recall were both 0.13. When considering indirect subsumption and direct but inverse relations as correct precision/recall reach 0.19/0.20. For selected examples the co-occurrences are sufficient to achieve results up to 0.38 for the two small GO branches “cellular_component” and “molecular_process”, as well as for “enzyme regulator activity”, a sub-branch of “biological_process”. For the whole branch “biological_process” the often complex terms could not be reliably identified in text. For selected sub-branches from MeSH a F-measure up to 0.60 was obtained. In gen-

eral, there are two reasons for the low overall results, apart from suitability of co-occurrence information for the task.

- The algorithm as implemented assigns at most one parent relationship. Both, GO and MeSH allow multiple parents per term. This may strongly affect performance in terms of recall.
- The co-occurrence measure we obtained on the basis of PubMed abstracts is document-wise or abstract-wise. The causal relationship for a co-occurrence of two terms is definitely influenced by the proximity of two terms. The strongest evidence will be provided by with-in sentence or next-sentence co-occurrence.

The evaluation results suggest that using life science literature is a possible approach for extending ontologies, although there still is much room for improvement of the evaluated approach. We think that co-occurrence information has the potential to aid taxonomy generation in applications to find relations where patterns do not occur. With growing corpora and better measures which possibly include distributional similarity measures used e.g. in Formal Concept Analysis (Ganter et al., 2005) we expect to achieve better results in the future.

5.6 Future Work

Syntactic information Once definitions containing taxonomic information are available for a term the mentioned potential parent term has to be mapped to some existing ontology term or a new term need to be proposed to be included in the ontology. This mapping currently only accepts perfect mentions and plural modifications. The mapping can be improved by adding better syntactic mapping relying on substitution patterns or external sources for synonyms. This way more relations to existing terms could be proposed.

Contextual information The extracted definitions are ambiguous for the term to be defined. Contextual information extracted from the ontology under construction and other previously accepted definitions will help to alter the ranking of definitions to prefer definitions for the intended sense of a term.

Reuse of existing ontologies Existing resources, such as the UMLS, BioPortal³ and the EBI ontology lookup service⁴ should be used to revise relations obtained by text-mining and should be used to add known relations defined in related ontologies.

Extension of the co-occurrence base method Currently the method evaluated in (Section 5.4) assigns at most one parent per term. Multiple inheritance is common to many ontologies and also to GO and MeSH we used in our evaluation. The method should be extended to allow the assignment of several parents per term.

The co-occurrences used to calculate the centrality network are document-wise co-occurrences. Alternatively the co-occurrence could be calculated on the level of sentences, paragraphs or sections if available. The influence of the scope or co-occurrence need to be further investigated to obtain meaningful relationships and to optimise the method to reliably prediction taxonomic relationships.

³ <http://www.bioontology.org/>

⁴ <http://www.ebi.ac.uk/ontology-lookup/>

Algorithms, Data Structures and Implementations

This chapter gives an overview over algorithms, data structures, and implementations for the ontology generation methods introduced in the Chapters 3, 4, and 5. The corresponding projects are listed in Table 6.1. For selected components of the once 3.1 and 4.1 the technical details will be provided. The components are either taggers or concept revisions. Taggers are used to annotated text. Concept revisions are used to manipulate ontology concept representations.

<i>Purpose</i>	<i>Project</i>	<i>Total</i>	<i>Main code</i>	<i>Test code</i>
Data structures for text annotation	ElivagarCore	10,369	7,995	2,374
Algorithms for text annotation	Elivagar	12,257	9,210	3,041
Datasources (Lucene, Database)	ElivagarDataSources	832	557	275
TNT POS Tagger, Java Integration	TNTWrapper	282	256	26
Term & definition generation	Idavoll	6,268	5,090	1,178
DOG4DAG Ontology Generation Tool	OBO-Edit	4,568	4,568	
Idavoll Term Generation Platform	IdavollPlatform	1,842	1,689	153
Analysis and Evaluation	IdavollAnalysis	10,180	9,688	492

Table 6.1. Overview over software projects implemented for text mining and ontology generation. In total the projects add up to 46,598 lines of code. Half of the code has been created in collaboration with others. Not listed are the projects with partially generated code, which are the web service and the client projects GoPubMedTermGenerationService, GoPubMedDefinitionGenerationService, GoPubMedOntologyLookupService, PubMedTokenStatisticsWebService, GoogleNgramService as well the Lucene indexing projects LuceneGoogleIndexing, LucenePubMedIndexing.

The most basic requirement for a text mining system is an efficient and convenient representation of text, which allows the specification of algorithms to markup functional or semantic text units. For the text mining and ontology generation projects, the we use an own implementation, the TextTree.

6.1 TextTree – a tree representation for text

A possible approach to structure text is to represent it as a connected acyclic simple graph. Such a data structure can easily be implemented as a DOM (Document Object Model). Several implementations of DOM trees are available. However, the available

DOM implementations are not optimised for text-mining. The Java implementation developed for this work specifically optimises the addressing of single characters and iterate forward and backward over nodes and axes, while the memory overhead is minimal. Figure 6.1 displays a tree structure segmenting a text into an abstract, sentences, and tokens. The tree representation combined with the index look-up for text positions as well as nodes by types allows fast annotation and look-up of text units tagged with certain types.

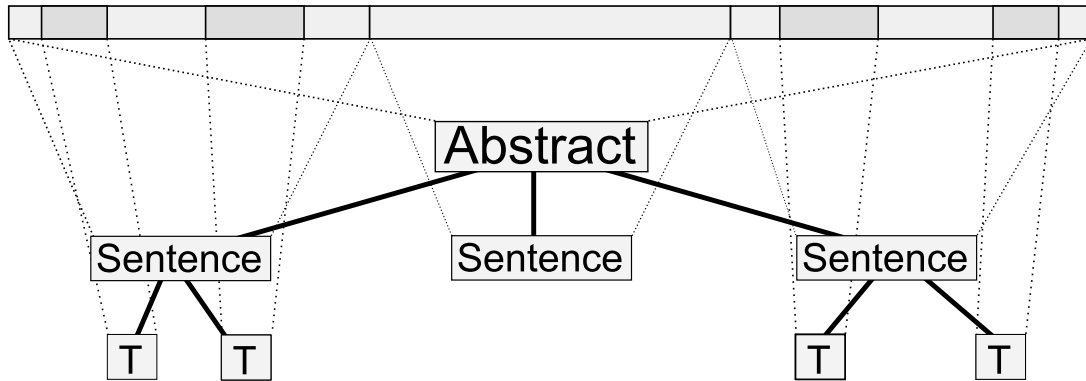


Fig. 6.1. The TextTree is a data structure used as representation for text. Each node in the tree represents a non-overlapping sub-string of the text. Nodes correspond to tagged regions, e.g. tokens, sentence. Nodes can hold several types of tags, e.g. tokens will have Part-of-Speech tags if they represent a word (Figure from Doms (2009)).

A text in the TextTree data structure is represented by a root node containing the whole text. By sequentially annotating this text the ranges for the non-overlapping annotations are positioned in the tree structure. To access all lexical units of the annotated text one can simply traverse all the leaf nodes. Because each node contains the start and end range, the text can be iterated at any level that has been previously annotated, e.g. sentences, tokens, noun phrases, abbreviations.

6.2 Taggers

The text manipulations and annotations are performed by taggers (implementations of `ElivagarTagger`, Listing 6.1). Taggers annotate text by inserting nodes or adding attributes to text nodes. Taggers insert new nodes by addressing a character range and adding a tag. Partially overlapping annotations are not allowed. While this clearly is a restriction in praxis it never occurred as a limitation. The following pseudo-code example illustrates the process of annotating:

```
text = "The murine embryonic stem cell test (EST) represents a validated alternative
       method for in vivo embryotoxicity testing."
text.annotate(0, 118, sentence)
text.annotate(37, 39, new Abbreviation("embryonic stem cell test", "EST"))
text.annotate(4, 34, noun_phrase)
```


Listing 6.1. ElivagarTagger.java

```

1 package elivagar;
2
3 /**
4  * Any Elivagar tagger takes a TextTree and processes
5  * it by modifying its data structure by splitting tree
6  * node into sub nodes or adding tags to existing nodes.
7  * The processing can return a result object of any kind
8  * for process supervision.
9  * All results of the processing needed in another tagger
10 * should be contained in the TreeText object.
11 */
12 public interface ElivagarTagger {
13
14     /**
15      * Does any processing on a TextTree object.
16      * @param text
17      * @return a result object containing status information
18      */
19     public TaggingResult process(TextTree text);
20 }

```

Tagger	Annotation	Dependencies
<i>ElivagarTokenizer</i>	lexical units: words, opening/closing brackets/quotes, punctuations, white spaces	
<i>SentenceTagger</i>	syntactical units: sentences	<i>ElivagarTokenizer</i>
<i>POSTagger</i>	Part-of-Speech syntactic categories: noun, verb, adverb, etc.	<i>SentenceTagger</i>
<i>NounPhraseTagger</i>	noun phrases, grouping of sequences of syntactic categories	<i>POSTagger</i>
<i>AbbreviationTagger</i>	finds local abbreviations	<i>ElivagarTokenizer</i>

Table 6.2. Selected implementations of the class ElivagarTagger

6.2.1 AbbreviationTagger

The task of finding abbreviations and lexical variants can be performed with good results (Section 2.3.3). For this work the method by Adar (2004) for finding local abbreviations has been implemented. As only a small domain relevant set of documents is used for term generation, there is no need for disambiguation. In cases where disambiguation is required the method of Gaudan et al. (2005) was selected for implementation.

6.2.2 PosTagger

Two types of Part-Of-Speech taggers have been used in the experiments. The LingPipeTagger using the Java implementation by Carpenter (2009) and the TNTTagger where I wrapped the TNT Tagger native implementation by Brants (2000) in a Java component. The differences between the two implementations regarding the extraction of terms was analysed in Section 3.5.1 (Dependency on the part-of-

speech tagger). The ranking of terms only changed marginally when exchanging taggers against each other.

6.2.3 NounPhraseTagger

The basis for annotating noun phrases is text. All tokens in the text have to be annotated with Part-Of-Speech tags before the noun phrases can be determined. For the two Part-Of-Speech implementations POSLingTagger and TNTTagger I grouped the Part-Of-Speech categories into the four classes adjective pattern (*adj*), noun pattern (*noun*), verb pattern (*verb*), and fill word pattern (*fill*). On this basis the noun phrase tagger extracts noun phrases of the form $[adj|verb] * [fill]\{2\}[noun]^+$, where *fill* are fill words like *of*, *the*, *for*, and others.

6.3 Concept revisions

Initial candidate concepts (implementation TextConcept) are created from the extracted noun phrases and abbreviations. Such a TextConcept has as a unique id, a label, the lexical representations found in text, as well as abbreviations, if available. Listing 6.3 shows the simple interface of all concept revision. A concept is in the first step accepted for revision and then processed. For instance, all ranking scores can only be calculated once the frequency of occurrences has been assigned for a concept. Therefore, only concepts with frequencies assigned are accepted for any of the revisions calculating ranking scores.

<i>Concept revisions</i>	<i>Annotation</i>	<i>Dependencies</i>
<i>MergeConceptRevision</i>	grouping concepts with all concepts in the transitive closure of the lexical representations of the concept	
<i>GlobalFrequencyRevision</i>	concepts are annotated with term and phrase frequencies obtained from PubMed and Google N-grams	
<i>ScoreCValueRevision</i>	calculation of the C-Value score like (Frantzi et al., 2000)	
<i>ScoreTfidfRevision</i>	calculation of tf-idf	<i>GlobalFrequencyRevision</i>
<i>ScoreHGRevision</i>	calculation of the conditional probability of occurrence for a concept given a large reference corpus	<i>GlobalFrequencyRevision</i>

Table 6.3. Selected implementations of the class ConceptRevision.

6.3.1 Merge Concept Representations

The grouping of lexical variants improves the term recognition result (Nenadić et al., 2004a). In DOG4DAG, all lexically overlapping concepts are also grouped. This grouping is not performed if concepts only have common abbreviations. Listing 6.2 shows the part of the implementation which performs this grouping.

Listing 6.2. Grouping of objects by overlap for the creation of the transitive closure of concepts over lexical representations in MergeConceptRevision.java

```

1  /**
2   * Add a group of element to be grouped
3   *
4   * @param ts
5   */
6  public void addIterable(Iterable<T> ts)
7  {
8      Set<T> temp = null;
9      Set<T> existing = null;
10
11     for (T t : ts) {
12         if (representativeToSet.containsKey(t)) {
13             // just found existing
14             existing = representativeToSet.get(t);
15             for (T element : ts) {
16                 existing.add(element);
17             }
18             if (null != temp) {
19                 // set existing as the set for temporary elements
20                 for (T tempElem : temp) {
21                     existing.add(tempElem);
22                 }
23                 for (T tempElem : temp) {
24                     representativeToSet.put(tempElem, existing);
25                 }
26             }
27             temp = existing;
28         }
29         else {
30             if (null == temp) {
31                 // no existing selected yet
32                 temp = new HashSet<T>();
33                 for (T element : ts) {
34                     temp.add(element);
35                 }
36             }
37             representativeToSet.put(t, temp);
38         }
39     }
40 }
41
42
43 /**
44  * Return the {@link Set} of elements for a representative
45  *
46  * @param object
47  * @return {@link Set} of T
48  */
49 public Set<T> getSetForRepresentative(T object)
50 {
51     return representativeToSet.get(object);
52 }
53
54 [...]

```

Listing 6.3. ConceptRevision.java

```

1 package elivagar.revision;
2
3 [...]
4
5 /**
6  * A strategy to derive zero or more new concepts from a given concept
7  */
8 public abstract class ConceptRevision
9 {
10     /**
11      * @param concept
12      * @return true if the revisor should be applied to the concept
13      */
14     public abstract Object accept(Concept concept);
15
16     /**
17      * @param concept
18      * @param meta information collected during acceptance
19      *           for this concept
20      * @return a (possibly empty) list of derived concepts
21      */
22     public abstract List<Concept> revise(Concept concept, Object meta);

```

6.3.2 Global Frequency Revisions

The global frequency revisions assign frequency counts to TextConcepts. Two sources for such statistics have been used. As domain independent source the content of web pages (indexed from Google) is used. As life sciences specific source all of PubMed was analysed.

Google frequencies (Google-N-Grams)

The web-derived frequencies have been obtained from the Web 1T 5-gram Version 1 (Brants and Franz, 2006), a resource of occurrence counts for 1-, 2-, 3-, 4-, and 5-grams which have been found in Google indexed web sites in 2005. The provided 24 GB compressed (gzipped) text files have been processed and a Lucene index, 150GB in size, has been created to access the data in the application. Tokenization in the Web 1T was performed in a way that

- hyphenated words were usually separated,
- hyphenated numbers usually form one token,
- sequences of numbers separated by slashes form one token, and
- URLs or email addresses are preserved as one token.

In statistics for the resource are shown in Table 6.4.

PubMed frequencies (PubMed-Cooc)

The approx. 18M scientific abstracts listed in PubMed have been sentences splitted and tokenized. For tokenization the ElivagarTokenizer was use, as it preserved

Number of tokens	1,024,908,267,229
Number of sentences	95,119,665,584
Number of unigrams	13,588,391
Number of bigrams	314,843,401
Number of trigrams	977,069,902
Number of fourgrams	1,313,818,354
Number of fivegrams	1,176,470,663

Table 6.4. Statistics on tokens, sentences and n-grams in the Google Web 1T 5-gram Version 1 (Brants and Franz, 2006).

Number of tokens	9,606,331
Number of pairwise co-occurrences (> 500)	2,132,265

Table 6.5. Statistics on the tokens and pairwise co-occurrences extracted from PubMed.

chemical formulas and resolves brackets. From all sentences the occurrence of single tokens and the pairwise co-occurrence has been calculated. In statistics for the resource are shown in Table 6.5.

6.3.3 Scoring Revisions

The scores obtained from the `ScoreCValueRevision` and the `ScoreTfidfRevision` can be calculated in linear time, the `ScoreHGRevision` is much more computational expensive. As example Table 6.6 shows scores `ScoreTfidfRevision` and `ScoreHGRevision` with frequencies obtained from Google indexed web sites and PubMed.

Scoring Revision: `ScoreHGRevision`

The hypergeometric distribution used to calculate scores for the `ScoreHGRevision` depends on 3 parameters

- the total number N of elements in the population
- the number $M \leq N$ of elements with a specific property contained in population
- the number $n \leq N$ of elements contained in the sample drawn

The probability distribution specifies the probability $P(X = k)$, that there exist k elements with the expected property in the sample.

$$h(k|N; M; n) := P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}},$$

with $\binom{N}{n}$ being the binomial coefficient. The probability that there exist at most or at least k elements with the property in the sample is described with the cumulative sum of probabilities.

$$H(k|N; M; n) := P(X \leq k) = \sum_{y=0}^k h(k|N; M; n) = \sum_{y=0}^k \frac{\binom{M}{y} \binom{N-M}{n-y}}{\binom{N}{n}},$$

The run-time complexity for the calculation of probabilities as defined in Section 3.5.3 (Hypergeometric distribution) is much higher than for e.g. tf-idf. The calculation of the binomial coefficients has a time complexity of $O(n^2)$ (Pascal's triangle

with $O(n^2)$ memory complexity). The probability is defined as the sum over the k elements having the property in the sample size. The probability has to be calculated for each word in the sample n . For large k this becomes very slow. Second, the obtained probabilities are usually very small ($< 10^{-10}$). When calculating the sum over n we stop if the probability does not change anymore. Also we worked with the exponent of the probabilities accepting the reduction of the accuracy and preventing number underflows.

Term	PubMed			Google		
	Global Frequency	Tf-Idf	PValue	Global Frequency	Tf-Idf	PValue
endosome	2,347	10.03	-81.92	51,656	8.21	-83.9
confocal	23,599	13.06	-58.83	410,068	9.9	-63.19
pancreas	90,949	15.85	-45.37	1,337,924	11.21	-51.39
microscopy	229,474	18.57	-36.19	3,154,576	12.4	-42.84
kidney	293,755	19.47	-33.76	7,514,651	13.9	-34.25
electron	263,828	19.07	-34.81	9,176,882	14.3	-32.28
liver	872,554	24.7	-23.18	12,139,938	14.89	-29.55
mouse	462,743	21.36	-29.3	33,668,364	17.56	-19.77
fish	97,125	16.02	-44.72	52,768,568	19.07	-15.66
transportation	8,378	11.5	-69.18	53,123,660	19.09	-15.6
protein	2,193,808	31.99	-14.66	53,260,540	19.1	-15.57
dog	83,877	15.65	-46.18	77,282,721	20.56	-12.33
heart	674,509	23.22	-25.65	90,275,312	21.24	-11.03
cell	2,787,789	34.64	-12.58	113,072,063	22.31	-9.23
all	1,717,193	29.66	-16.86	2,022,464,594	62.57	-2.75
endosome	2,347	10.03	-81.92	51,656	8.21	-83.9
transportation	8,378	11.5	-69.18	53,123,660	19.09	-15.6
confocal	23,599	13.06	-58.83	410,068	9.9	-63.19
dog	83,877	15.65	-46.18	77,282,721	20.56	-12.33
pancreas	90,949	15.85	-45.37	1,337,924	11.21	-51.39
fish	97,125	16.02	-44.72	52,768,568	19.07	-15.66
microscopy	229,474	18.57	-36.19	3,154,576	12.4	-42.84
electron	263,828	19.07	-34.81	9,176,882	14.3	-32.28
kidney	293,755	19.47	-33.76	7,514,651	13.9	-34.25
mouse	462,743	21.36	-29.3	33,668,364	17.56	-19.77
heart	674,509	23.22	-25.65	90,275,312	21.24	-11.03
liver	872,554	24.7	-23.18	12,139,938	14.89	-29.55
all	1,717,193	29.66	-16.86	2,022,464,594	62.57	-2.75
protein	2,193,808	31.99	-14.66	53,260,540	19.1	-15.57
cell	2,787,789	34.64	-12.58	113,072,063	22.31	-9.23

Table 6.6. Examples for scoring revisions with Tf-Idf and PValue for PubMed and Google statistics
The Tf-Idf and probability scores (PValue) are calculated using frequencies from PubMed and n-grams frequencies from the Google Web 1T 5-gram Version 1 corpus (Brants and Franz, 2006). The first part is sorted by Google-based PValue scores, the second part by PubMed based PValue scores. Terms like “transportation” and “dog” change rank significantly both are rare in PubMed and frequent in the web. Terms like “endosome” and “cell” are ranked similar, “cell” is frequent, “endosome” is rare in PubMed and in the web. For all terms Tf-idf lead to the same ranking. While the PValue is normalised between PubMed and Google and can be combined in one score, Tf-Idf is not.

6.3.4 Contributions to implemented software

Many of the software components developed for the implementation of the ontology generation methods have been shared projects. In the following, the participating developers are listed for shared project.

Ontology Learning

- *Idavoll* – algorithms for extracting and ranking terms and definitions
- *IdavollPlatform* – web application build on Google GWT to access to demo the term generation methods

General text mining data structures

Shared work with Loic A. Royer and Andreas Doms. Text-mining data structures and general Taggers (Tokenizers, Stemmers, etc)

- *ElivagarCore* – data structures and annotation framework
- *Elivagar* – general text-mining and word sense disambiguation

Ontology learning web services

Web services to provide access to ontology learning methods for the ontology editors OBO-Edit, Protégé and in GoPubMed.

- *GoPubMedOntologyGenerationServiceLogModule*
- *GoPubMedDefinitionGenerationService*
- *GoPubMedOntologyLookupService*
- *GoPubMedTermGenerationService*

Resource web services

- *GoogleNgramService* – web service to provide a cached access to an index over the large of WebCT n-grams source.
- *PubMedNGramWebService* – web service to provide access to n-grams extracted from 18 million PubMed abstracts
- *PubMedTokenStatisticsWebService* – web service to access token frequencies and sentence-wise co-occurrences extracted from 18 million PubMed abstracts

Programmatic access to GoPubMed

Software to access GoPubMed documents and annotations

- *GoMeshPubMed* – access all documents and annotations like in GoPubMed
- *PubMedSearch* – provide search in PubMed
- *PubMedSearchViaYggdrasil* – provide search like in GoPubMed caches
- *YggOntologies* – access the ontologies used in GoPubMed

Lucene Indexing

The fulltext indexing of PubMed is shared work with Heiko Dietze.

- *LuceneGoogleIndexing* – indexing all *n*-grams contained in the Google WebCT corpus
- *LucenePubMedIndexing* – indexing PubMed abstract
- *ElivagarDatasourcesLucene* – framework adapter to access Lucene indices
- *PubMedFullTextIndex* – indexing and accessing an Lucene PubMed fullext index

Other software component

- *MSNLiveSearchClient* – Microsoft Live Search client
- *ElivagarVisualization* – visualizing annotated text trees (shared work with Loic A. Royer)

Integration of Ontology Generation in Ontology Editors

References

Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. In *ISMB (Supplement to Bioinformatics)*, Impact factor 2009: 4.3 (accepted for publication)

Winnenburg, R., Wächter, T., Plake, C., Andreas, D., and Schroeder, M. (2008). Facts from text: Can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Briefings in Bioinformatics*, 9(6):466–478, Impact factor 2009: 4.6)

Conferences / Workshops

Wächter, T. and Schroeder, M. (2009). An Ontology Generation Plugin for OBO-Edit. *3rd International Biocuration Conference*, April 16-19, Berlin, Germany

Wächter, T. (2009). The Ontology Generation Tool for OBO-Edit. *Presentation at the GO Consortium & SAB Meeting*, September 23-25, Cambridge, United Kingdom

The DOG4DAG ontology generation methods developed in this thesis have been seamlessly integrated in OBO-Edit and Protégé, two widely used ontology editors in the life sciences. The systems offers either to submit a query to PubMed or the Web or to upload text or PDF documents. While PubMed is the default source for terminology, the Web is often useful since full-text articles and other on-line resource can be implicitly included in the search. When adding a term to the ontology, possible parents are suggested on the basis of generated definitions. Existing terms from other ontologies are automatically cross-referenced.

It has been shown on recent examples (Section 7.6) how the OBO-Edit Ontology Generation Tool can support the annotation of genes and gene products and the associated extension of the Gene Ontology.

In addition, a novel collaborative taxonomy editor has been specified as user-friendly, web-based alternative to existing ontology editors. It allows domain experts to contribute to the Go3R ontology without having to install or learn new complex software systems. The editor directly modifies the ontology of the semantic search engine which immediately has effect on subsequent searches.

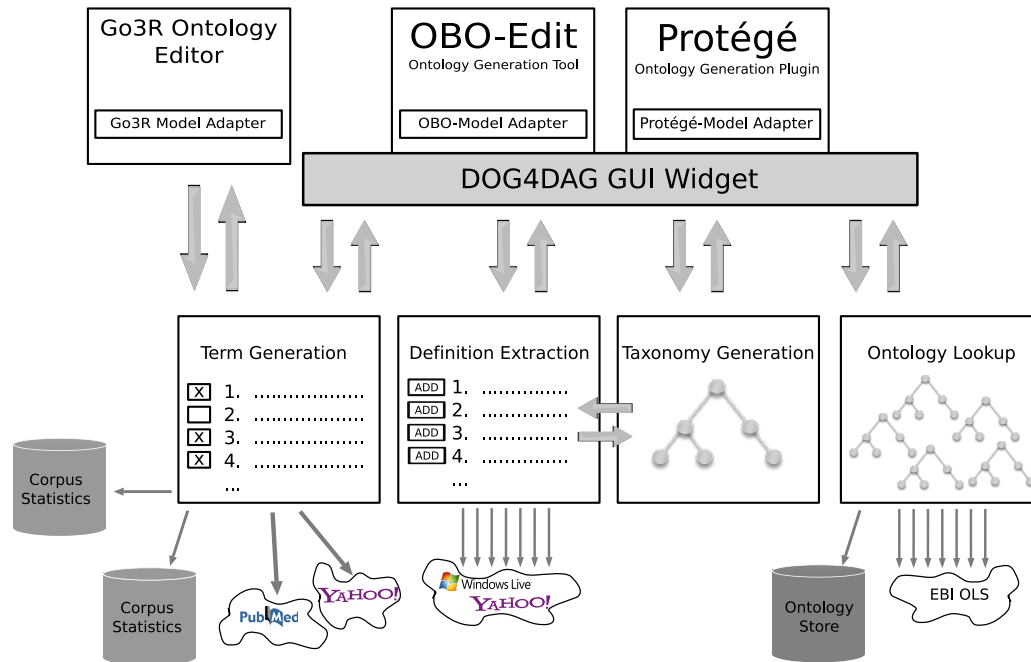


Fig. 7.1. Overview on the integration of ontology learning methods in ontology editors. The term generation use several corpus statistics for frequencies and co-occurrences of terms, retrieves documents from PubMed and the Yahoo search engine. The definition generation uses beside the Yahoo the Windows Live Search engine. The ontology look-up is performed using the Ontology Look-up Service provided by the European Bioinformatics Institute, Cambridge, UK.

7.1 Introduction

As the scientific truth advances, ontological knowledge needs to evolve. Ontologies need to be maintained. This evolution process includes adding new concepts, the deletion of obsolete concepts, re-structuring of already defined concepts as well as adding synonyms, definitions, and relations. Creating and maintaining such ontologies is a labour intensive, difficult, manual process.

In previous chapters it has been evaluated to what extent semi-automatic ontology generation methods can support this process. In order to contribute to automation of ontology generation, algorithms and methods as developed in this thesis have been integrated into Protégé and OBO-Edit, two widely used editors in the life sciences.

Figure 7.1 provides an structural overview for the presented software in this chapter. All three editors share the same service infrastructure for term generation, definition extraction, taxonomy generation, and ontology look-up. OBO-Edit and Protégé share the DOG4DAG GUI widget which encapsulates all ontology generation functionality and the communication to the web services. For each editor specialised adapters had to be implemented for the different ontology models and the plug-in mechanisms.

OBO-Edit Ontologies and taxonomies have proven highly beneficial for biocuration. The Open Biomedical Ontology Foundry (www.obofoundry.org) alone lists over 90 ontologies mainly built with OBO-Edit. To address the needs of biocurators ontology generation methods have been integrated in OBO-Edit the ontology editor developed and maintained by the Gene Ontology Consortium.

Protégé To give equally support to developers of ontologies in OWL, the term and definition generation has been integrated in Protégé, a widely used ontology editor.

Go3R Editor Ontology development, as performed in this thesis, is also largely motivated by the application of ontology-based literature search. The major bottleneck here is the availability of suitable ontologies. To be able to transfer the technology to other knowledge domains new ontologies need to be created. A review on existing editors revealed, that none of the existing tools meet the requirement for the collaborative creation of taxonomies. To overcome this limitation a novel ontology editor has been specified to support collaborative ontology development and integrate ontology generation methods.

Outline

Following a brief overview on existing ontology editors (Section 7.4), in this chapter a detailed description of the OBO-Edit Ontology Generation Tool is provided. It will be explained on real tasks performed by researchers editing the Gene Ontology, how this new tool can be used in the process of annotating genes and proteins as well as for the resulting extension of the Gene Ontology (Section 7.6).

Secondly, a novel user-friendly editor for taxonomies as the one used by the semantic search engine Go3R is introduced.

- Integration of ontology learning methods in the widely recognized editors OBO-Edit (Section 7.2) and Protégé (Section 7.3)
- Design and development of an ontology editor for Go3R and the integration of ontology learning methods (Section 7.4)
- Introduction to the web-based term generation platform (Section 7.5)
- Summary and Discussion (Section 7.7)
- Future Work (Section 7.8)
- Contributions (Section 7.9)

7.2 OBO-Edit Ontology Generation Tool

In the initial motivation (Section 2.1) the current situation of ontology development in the life sciences and the need for tailor-made support for biocuration was described. Beside the collection, annotation, and validation of raw experimental data (Bourne and McEntyre, 2006), biocurators have also developed a significant number of ontologies they use i.e to consistently annotate genes and proteins from different model organism (Ashburner et al., 2000). These ontologies are still under development and need to be maintained. We will show on examples how the ontology generation methods developed and evaluated in this thesis can support biocuration and the development of biological and biomedical ontologies. Term generation (2.3.1), definition generation (2.3.4), and taxonomy generation (2.3.5) methods can only provide this support when they are integrated in editors used by the community. Therefore the OBO-Edit Ontology Generation Tool has been developed (Wächter and Schroeder, 2010).

OBO Edit is an editor optimised for the development of ontologies in OBO-Format¹ which are frequently employed and created in the biocuration community. The editor supports several views onto the ontology model. The *Ontology Tree Editor* allows to browse and edit the taxonomic structure of the ontology like a directory tree. A *Graph Editor* allows to see and edit terms as they are embedded in the ontology graph. The *Text Editor* displays all information known for a term. For quality assurance, OBO-Edit contains build-in validation. Whenever the ontology is altered validators can be configured to revise the changes, which include:

- spelling checks on comments, definitions, names, and synonyms
- name redundancy checks
- dbxref checks to ensure references have been provided for definitions and synonyms
- namespace checks to ensure that each term has the correct namespace (e.g. one of the Gene Ontology parts *biological_process*, *cellular_component*, or *molecular_function*)
- checks on is_a completeness: check that every term has an all-is_a path to the root node

7.2.1 Ontology generation in three steps.

In a three step procedure the OBO-Edit Ontology Generation Tool supports the creation of new ontology terms from text. Free text, a query for PubMed abstracts, a web search query, or PDF documents can be used as source for terms. In a first step the terminology mentioned in the text is retrieved and ranked according to its importance in the domain. Abbreviations and lexical variants are recognised and terms similar to existing OBO terms are indicated as displayed. The list of candidate terms can be searched and filtered with regular expression patterns to focus on specific lexical aspects of interest. In a second step definitions for terms are generated and presented to the curator. The defined candidate terms, which are enriched with

¹ OBO Format Guide is available online under http://www.geneontology.org/G0.format.obo-1_2.shtml

All Steps Terms Generation Definition Generation Help About

Step 1: Term Generation

PubMed Web Text PDF Clipboard (12)

Query PubMed:

Terms generated for PubMed query 'lipoprotein'			
<input checked="" type="checkbox"/>	lipoprotein [FMA:63169, CHEBI:6495]	DEF	1
<input checked="" type="checkbox"/>	cholesterol [IMR:0200486, CHEBI:16113]	DEF	1
<input checked="" type="checkbox"/>	low-density lipoprotein (generated: 'low density lipoprotein') [FMA:63170, CHEBI:39026, GO:0005322]	DEF	1
<input checked="" type="checkbox"/>	high density lipoprotein [GO:0005321, CHEBI:39025, CCO:F0001556]	DEF	1
<input checked="" type="checkbox"/>	lipid [MOD:00390, MOD:00668, MOD:00392, CHEBI:18059, IMR:0001362, FMA:67264]	DEF	1
<input checked="" type="checkbox"/>	triglyceride [CHEBI:17855]	DEF	1
<input type="checkbox"/>	level	DEF	1
<input checked="" type="checkbox"/>	risk	DEF	1
<input type="checkbox"/>	low-density [PATO:0001790]	DEF	1
<input checked="" type="checkbox"/>	low density lipoprotein cholesterol [CHEBI:47774]	DEF	1
<input checked="" type="checkbox"/>	oxidized low density lipoprotein	DEF	1
<input type="checkbox"/>	high density lipoprotein cholesterol [CHEBI:47775]	DEF	1
<input checked="" type="checkbox"/>	apolipoprotein [CCO:F0001556, GO:0005319, GO:0005320, CHEBI:39015]	DEF	1
<input type="checkbox"/>	metabolic syndrome [DOID:14221, EFO_0000195]	DEF	1

Search: Filter: ☐ Show existing terms only.

Step 2: Definition Generation

Find Definitions for:

Definitions for "lipoprotein"			
+	lipoprotein is a combination of fat (lipid) and protein that wraps around the individual fat molecules in our bodies,	DEF	1
+	lipoprotein is a biochemical assembly that contains both protein s and lipid s and may be structural or catalytic in function.	DEF	1
+	Lipoprotein is an assembly of proteins and lipids that carry cholesterol between the liver and the body tissues.	DEF	1
+	lipoprotein is a round particle that has either certain types of cholesterol or triglycerides at its center.	DEF	1
+	lipoprotein is a substance in the blood that carries cholesterol and other fats to the body's cells.	DEF	1
+	Lipoprotein is a protein molecule that transports fat or "lipid" around in your bloodstream.	DEF	1
+	lipoprotein is an obligatory factor for experimental hypertriglyceridemia in nephrotic rats.	DEF	1
+	lipoprotein is a biochemical assembly that contains both proteins and lipids and may be structural or catalytic in function.	DEF	1
+	lipoprotein is a biochemical assembly that contains both Proteins and Lipids The lipids or their derivatives may be	DEF	1
+	Lipoprotein is a predominant Toll-like receptor 2 ligand in Staphylococcus aureus cell wall components Masahito	DEF	1
+	Lipoprotein is a biochemical assembly that contains both proteins and lipids.	DEF	1
+	Lipoprotein is an assembly of proteins and lipids that carry cholesterol...	DEF	1
+	Lipoprotein is an important subject that is taught at all academic levels.	DEF	1
+	Lipoprotein is a chemical compound that contains both proteins and lipids.	DEF	1

Filter:

Edit Definition

A biochemical assembly that contains both proteins and lipids and may be structural or catalytic in function. Lipoproteins carry cholesterol between the liver and the body tissues.

Abbreviations, synonyms and known children

☐ LP (Abbreviation)

☐ LPs (Abbreviation)

☒ low-density lipoprotein...

☐ Pulmonary surfactant (is...)

Step 3: Add to Ontology

Add Term: ☒ Include children

Potential parent terms (existing in OBO-Edit):

Selec...	Predicted	Relation	Term	Comment
<input type="checkbox"/>	identical	is_a	lipoprotein (ID:0000000)	same as existing ter...
<input checked="" type="checkbox"/>	is_a	is_a	protein (ID:0000008)	validated, predicted
<input type="checkbox"/>	sub_class_of	is_a	high density lipoprotein (ID:0000003)	predicted, similar term
<input type="checkbox"/>		is_a	low-density lipoprotein (ID:0000002)	similar term
<input type="checkbox"/>		is_a	very low density lipoprotein (ID:0000006)	similar term
<input type="checkbox"/>		is_a	oxidized low density lipoprotein (ID:0000004)	similar term
<input type="checkbox"/>		is_a	intermediate-density lipoprotein (ID:0000005)	similar term
<input type="checkbox"/>		is_a	is_a (OBO_REL:is_a)	existing term
<input type="checkbox"/>		is_a	Myelin (ID:0000009)	existing term
<input type="checkbox"/>		is_a	disease (ID:0000001)	existing term
<input type="checkbox"/>		is_a	part_of (part_of)	existing term
<input type="checkbox"/>		is_a	obo:TERM (obo:TERM)	existing term
<input type="checkbox"/>		is_a	obo:TYPE (obo:TYPE)	existing term
<input type="checkbox"/>		is_a	obo:link (obo:link)	existing term
<input type="checkbox"/>		is_a	union of (union_of)	existing term

Filter: ☐ Show ticked parent terms only.

Fig. 7.2. Overview over the OBO-Edit-Ontology Generation Tool.

synonyms are in the final step inserted into the ontology. We addressed the difficulty of finding the correct position in a tree structure by using all information available to the plug-in. All potential parent terms e.g. all terms of the Gene Ontology, are displayed as list and are ranked higher, if they are (a) selected in OBO-Edit, (b) have a certain lexical overlap with the new candidate concept, (c) are contained in the specified definition, or (d) evidence for an relationship could be found in any of the OBO listed ontologies. With the novel Ontology Generation plug-in for OBO-Edit 2, a contribution to the community is made by increasing the tool support for the development and maintenance of biomedical ontologies. In the following sections each generation step is described separately.

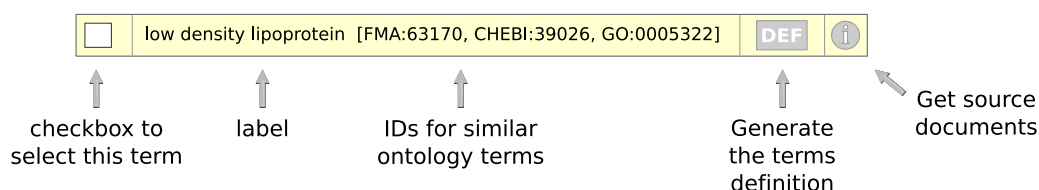
Step 1: Term Generation

For the easy to use integration of the DOG4DAG term generation method in OBO-Edit the following requirements have been addressed:

Requirements

1. **List representation of selectable, ranked term candidates:**
Users should be able to quickly verify the most prominent terms in a document set. Therefore each generated candidate term should be displayed in a way, that user input can be provided to validate or discard generated terms.
2. **Ontology look-up:** Indication of existing ontology terms, both within the ontology under construction as well as in external resources. Users should be informed that a term already exist in “some” ontology, hence was judged by another human to have the potential to be a term.
3. **Filtering of terms lists:** Filtering with syntactic criteria to show terms following a pattern or terms which exist in one or another form.
4. **Abbreviation detection:** Abbreviations mentioned in the texts used for extracting terminology should be identified and terminology and abbreviation need to be grouped to one concept.
5. **Textual sources:** Commonly used textual sources need to be supported. This includes PubMed and web content as well as locally available text.

The Term Generation view within the Ontology Generation Tool (Figure 7.3) presents the ranked list of terms (**Requirement 1**). Behind each term the identifiers of similar existing terms contained in the Open Biomedical Ontologies are listed. Terms are regarded as similar if they share a label or some variation of it (**Requirement 2**). The EBI Ontology Look-up Service is used to retrieve the identifiers and known children for all displayed terms. On the right side of each term a button to trigger the generation of definitions for this term is followed by an information button to retrieve the document set the terms have been extracted from.



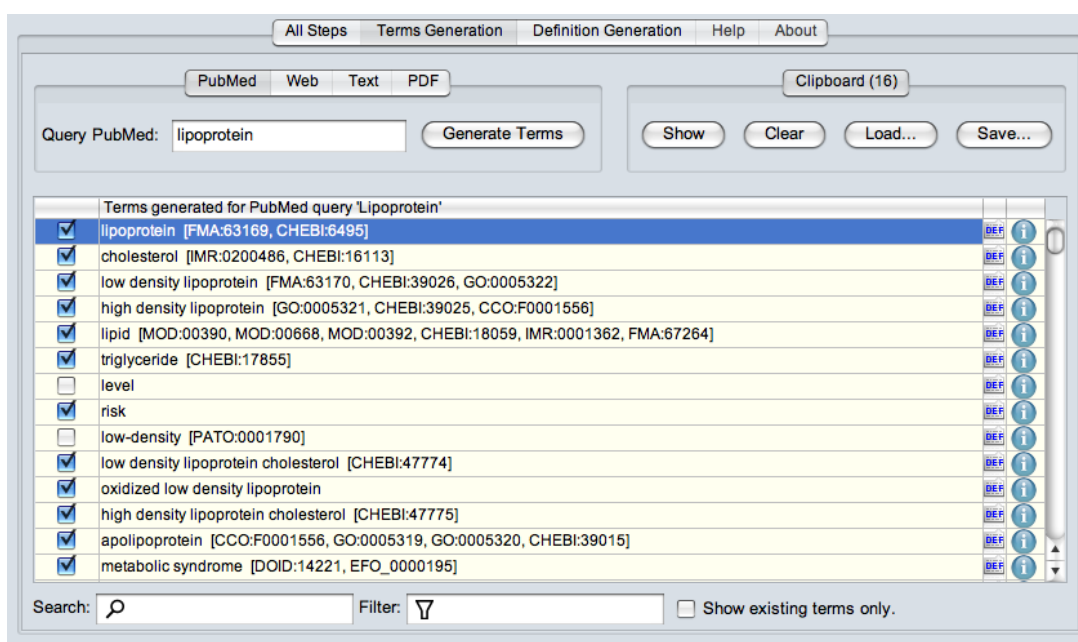


Fig. 7.3. OBO-Edit Ontology Generation Tool: Term generation view. Displayed are the terms extracted from PubMed abstracts containing the word “lipoprotein”. Alternatively, the short summaries typically provided by web search engine (snippets), PDF documents, text can be used. Each term can be selected and added to the clipboard. Examples shows terms from the Lipoprotein Metabolism Benchmark (Section 3.3).

The candidate terms are extracted from text. This text can currently be acquired from four different sources visible as tabs above the query field (**Requirement 5**).

- **PubMed:** the submitted query is sent to PubMed and the top 250 abstracts are being analysed
- **Web:** the submitted query is sent to a web search engine, currently Yahoo, and the top 1000 snippets (short summaries) are being analysed
- **Text:** the pasted text is being analysed
- **PDF:** the single PDF, or all PDF's in the specified folder are parsed and the contained text is being analysed

Abbreviations are being extracted and displayed with each term (**Requirement 4**). The candidate term list can be filtered to show only terms following a certain pattern, e.g. ends with lipoprotein (Figure 7.4). Regular expressions are supported². With the option “Show existing terms only”, the candidate term list can be filtered (**Requirement 3**) to gain a quick overview which terms of the ontology currently build with OBO-Edit are also contained in the analysed text (Figure 7.5). Generally, such terms are displayed in **bold face**. Terms can also be searched and filtered using regular expressions.

² see <http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html>

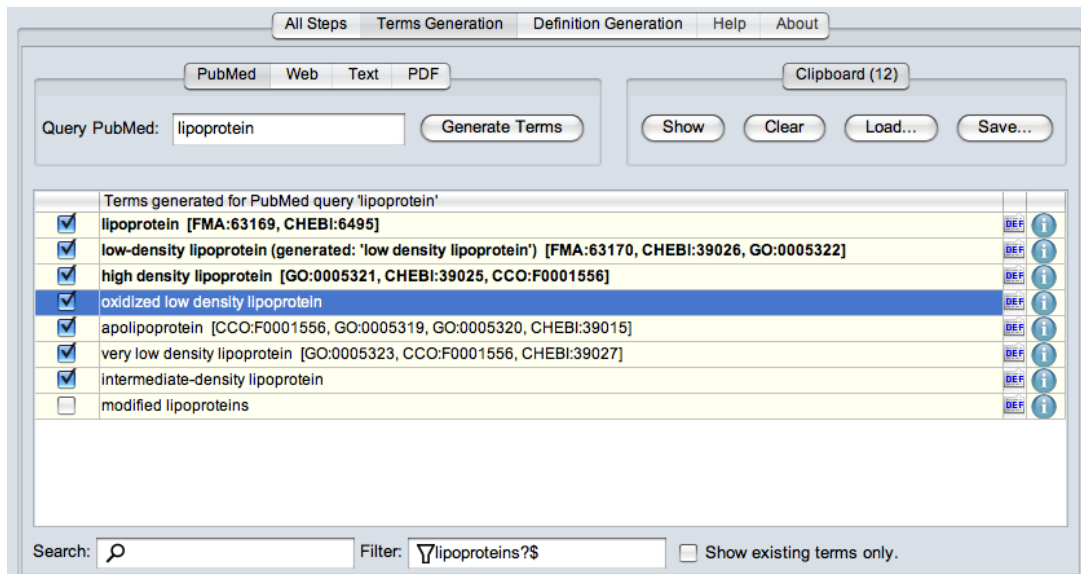


Fig. 7.4. OBO-Edit Ontology Generation Tool: Filtering by pattern in the term generation view. Generated terms like in Figure 7.3, but the term list has been filtered to show only terms ending with the word lipoprotein or lipoproteins. Regular expressions are supported. Terms that exist in the ontology loaded in OBO-Edit are displayed in **bold face**.

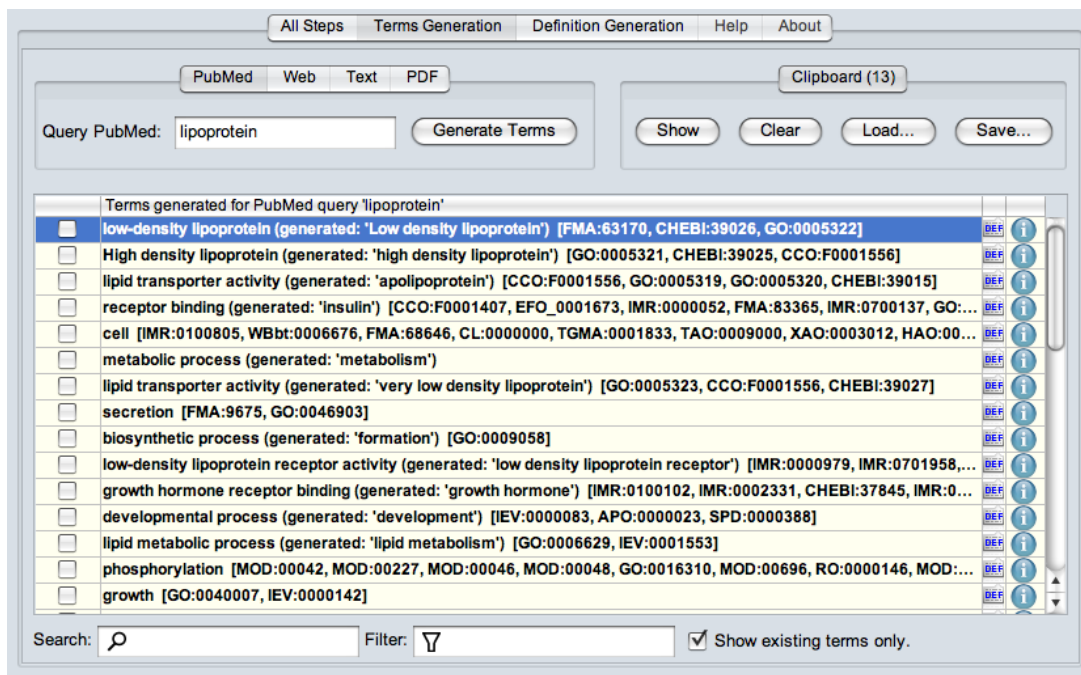


Fig. 7.5. OBO-Edit Ontology Generation Tool: Filtering by existing ontology term in the term generation view. Generated terms like in Figure 7.3, but restricted to existing terms loaded in OBO-Edit, in this case terms from the Gene Ontology as of 2009. Terms that exist in the ontology loaded in OBO-Edit are displayed in **bold face**.

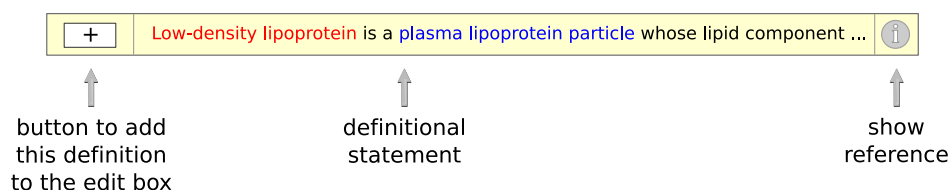
Step 2: Definition Generation

In Chapter 4 (Definition Extraction) general requirements have been formulated for the task of definition generation. This list can be extended with specific requirements for an application to generate definitions for terms on the basis of text references.

Requirements

1. **Domain independence:** The method in general should be **domain independent** to allow the creation of definitions for terms and concepts from diverse knowledge domains.
2. **Run-time requirement:** The method should be fast to allow **On-The-Fly interactive generation** of definition candidates.
3. **Filtering:** Like for terms, filtering of definitions is required to efficiently select the definitions of interest.
4. **Highlighting:** For better visibility the term to be defined (definiendum) and the head noun phrase of the definitional facts (definiens) should be **highlighted**.
5. **Integration:** As there might be terms generated for which there exist a term in the ontology loaded in the ontology editor, the term generation tool need to be aware of this and display existing properties appropriately.

The definition generation view (Figure 7.6) presents a list of generated definitions. The definitions are collected from web search results which are available independent from the domain (**Requirement 1**). Short summaries (snippets) retrieved from web search engines are frequently truncated within sentences. To reduce the processing time (**Requirement 2**) only the definitions visible to the user are being completed by consulting the original web resources and extracting the full sentences. The generated definitions are ranked as described in Section 4.1 (DOG4DAG Definition Extraction Method). Regular expression filtering (**Requirement 3**) allows an easy search for specific mentions in the generated definitions. The defined term (red), the definitional pattern (italic), and the head noun phrase of the differentia (blue) are marked up (**Requirement 4**). By clicking on the button with the “+” the definition is added to the edit box below the list of definitions for further editing. On the right side of the edit box the abbreviations and known children for the term are displayed for selection. The URLs to the sources of a definition on the web are accessible via the information button “show reference” behind each generated definition (Figure 7.8). Whenever terms from the ontology loaded in OBO-Edit can be mapped to a generated term, the existing definition is displayed (**Requirement 5**).



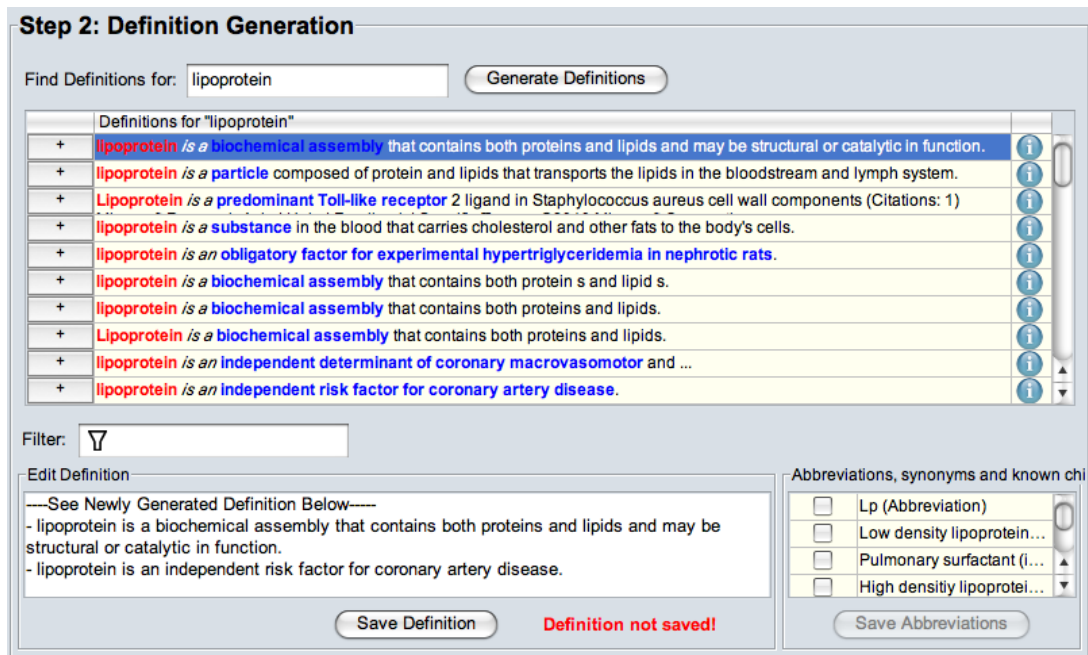


Fig. 7.6. OBO-Edit Ontology Generation Tool: Definition generation view. The generated definitions are ranked and the defined term, the definitional pattern, and the head noun phrase of the differentia are highlighted. Two definitions generated for “lipoprotein” have been added to the “Edit Definition” text field. Abbreviations, synonyms and known child terms from other ontologies are listed in the bottom right.

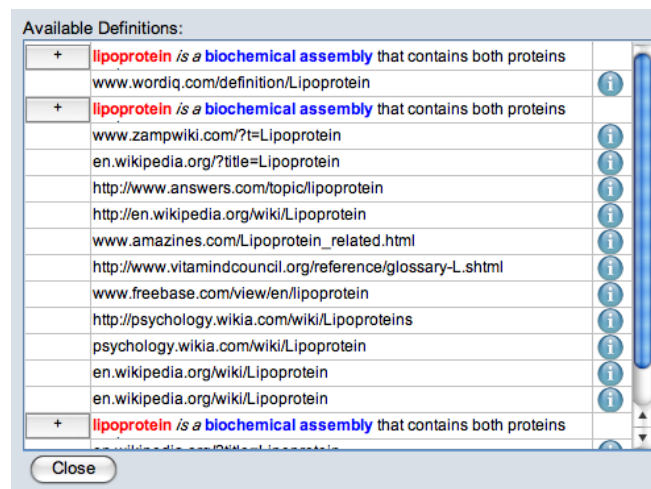


Fig. 7.7. OBO-Edit Ontology Generation Tool: reference information for generated definitions. The URLs of the web resources a definition has been extracted from are accessible via the information button behind each generated definition.

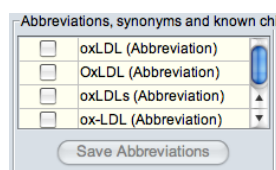


Fig. 7.8. OBO-Edit Ontology Generation Tool: abbreviations view. For each term the abbreviations found in the analysed are displayed, here for “oxidized low density lipoprotein”.

Step 3: Add To Ontology

In the last step, generated terms can be added to the ontology loaded in OBO-Edit. The DOG4DAG taxonomy generation method (Chapter 5) has been implemented in OBO-Edit. The following requirements have been addressed.

Requirements

1. **Filtering:** Without additional information, all terms defined in the ontology model of the ontology editor are potential parent terms. Filtering and searching in an alphabetically sorted list of terms is the simplest and most intuitive representation of the candidate parent terms.
2. **Selecting terms:** List selection is the simplest way to select parent terms from the ontology. Known parents can be automatically selected. High confidence predictions can be automatically selected.
3. **Add to ontology:** All predicted parent terms for a term should be added to the ontology in one operation.

The “Add to Ontology” view (Figure 7.10) shows a ranked list of all terms existing in the ontology loaded in OBO-Edit. Terms are ranked higher, if they are selected in OBOEdit, have a significant lexical overlap with the new candidate concept, are contained in the specified definition, or evidence for an relationship could be found in any of the OBO listed ontologies. Filtering the term list provides easy access to all potential parents (**Requirement 1**). Multiple parents can be selected (**Requirement 2**). The relationship type can be manually changed. All new parent terms are added in on revertible change transaction (**Requirement 3**). With the option “include children” enabled in Figure 7.10 add all known child terms found in other exiting ontologies are added including the reference to external existing ontology terms and the URL to the source of the accepted definitions (Figure 7.11).

Five potential parent terms from MeSH were suggested for the term “*Apolipoproteins*”: “*Proteins*”, “*Carrier Proteins*”, “*Family*”, “*Ligands*”, and “*Glycoproteins*”. The basis for these suggestions are the generated definitions listed in Table 7.1. The relation to “*Protein*” is correct and already indirectly exists (Figure 7.9). The term “*Family*” has been correctly found in the definition, but a different sense, not the sense for protein family, has been suggested. “*Ligands*”, “*Carrier Proteins*”, and “*Glycoproteins*” are correct prediction.

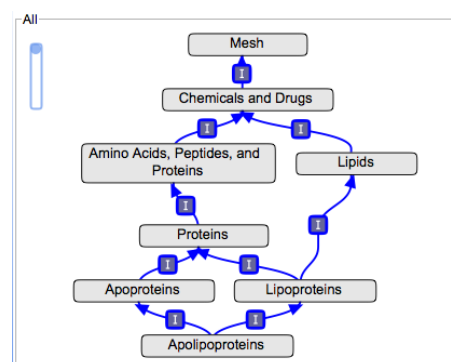


Fig. 7.9. OBO-Edit graph viewer. Ancestors of the term “*apolipoprotein*”.

Correct Term	Definition
(✓) Proteins	apolipoprotein is a hydrophobic 12-kDa protein processed from ...
✓ Carrier Proteins	Apolipoproteins are <u>carrier proteins</u> that combine with lipids to form ...
Family	apolipoproteins are <u>family</u> of amphipathic lipoproteins that plays a ...
✓ Ligands	apolipoproteins are <u>ligands</u> for hepatic lipoprotein receptors, ...
✓ Glycoproteins	apolipoprotein is a <u>plasma glycoprotein</u> involved in ...

Table 7.1. Listing of the source definitions for the predicted taxonomic relations for the term “apolipoprotein”. All suggested parents for “apolipoprotein” have been correctly extracted from the generated definitions. The relation to “protein” indirectly exists over the parent “apoprotein”.

Step 3: Add to Ontology

Add Term: ☐ Include children

Potential parent terms (existing in OBO-Edit):

Selec...	Predicted	Relation	Term	Comment
<input type="checkbox"/>	identical	is_a	Apolipoproteins (1053)	same as existing term
<input checked="" type="checkbox"/>	is_a	is_a	Apoproteins (1059)	validated
<input checked="" type="checkbox"/>	is_a	is_a	Lipoproteins (8074)	validated
<input type="checkbox"/>	sub_class_of	is_a	Proteins (11506)	predicted
<input type="checkbox"/>	sub_class_of	is_a	Carrier Proteins (2352)	predicted
<input type="checkbox"/>	sub_class_of	is_a	Glycoproteins (6023)	predicted
<input type="checkbox"/>	sub_class_of	is_a	Ligands (8024)	predicted
<input type="checkbox"/>	sub_class_of	is_a	Family (5190)	predicted
<input type="checkbox"/>		is_a	Id (7060)	existing term
<input type="checkbox"/>		is_a	Air (388)	existing term
<input type="checkbox"/>		is_a	Arm (1132)	existing term
<input type="checkbox"/>		is_a	DDT (3634)	existing term
<input type="checkbox"/>		is_a	DNA (4247)	existing term
<input type="checkbox"/>		is_a	Ear (4423)	existing term
<input type="checkbox"/>		is_a	Ego (4532)	existing term

Filter: ☐ Show ticked parent terms only.

Add term to Ontology

Fig. 7.10. OBO-Edit Ontology Generation Tool: Add to ontology view. For the term “Apolipoproteins” the five potential parent terms “Proteins”, “Carrier Proteins”, “Family”, “Ligands”, and “Glycoproteins” from the Medical Subject Headings were suggested on the basis of generated definitions. The two existing parents are shown selected.

Text Editor

ID: 0000001
Namespace: default_namespace
Name: lipoprotein

Definition | Comment | Cross Products

Definition: Lipoprotein is a biochemical assembly that contains both proteins and lipids and may be structural or catalytic in function. It transports the lipids in the bloodstream and lymph system and is an independent risk factor for coronary artery disease.

Dbxrefs: [URL:www.wordiq.com/definition/Lipoprotein](http://www.wordiq.com/definition/Lipoprotein)
[URL:http://www.wordiq.com/definition/Lipoprotein](http://www.wordiq.com/definition/Lipoprotein)
[URL:www.exampleproblems.com/wiki/index.php/Lipoprotein](http://www.exampleproblems.com/wiki/index.php/Lipoprotein)
[URL:http://www.cvd.idf.org/Risk_Factors/Diabetes__A_Major_Risk_Fac](http://www.cvd.idf.org/Risk_Factors/Diabetes__A_Major_Risk_Fac)
[URL:advancedcholesterolmanagement.com/images/lipidsSegrest.pdf](http://advancedcholesterolmanagement.com/images/lipidsSegrest.pdf)

Dbxrefs | **Synonyms** | **Subsets**

Lipoproteins
Scope: Related Synonym

Lipoprotein
Scope: Related Synonym

Xrefs: REF_OBO:FMA:63169

Fig. 7.11. Text editor view in OBO-Edit showing the attributes of a generated term. The term “lipoprotein” has been generated, defined and added to the ontology. The references (Dbxref) to the source documents for the definition were automatically added. Synonym entries with reference to existing ontologies, here the anatomy ontology FMA, were created.

7.3 Protégé Ontology Generation Plug-in

While the biocuration community largely works with OBO-Edit, the ontology editor Protégé is widely used for the creation of medical ontologies and whenever ontologies represented in the Web Ontology Language OWL are created. The DOG4DAG ontology generation methods have been integrated in Protégé 4 in the same way as done for OBO-Edit. A screen shot is shown in Figure 7.12.

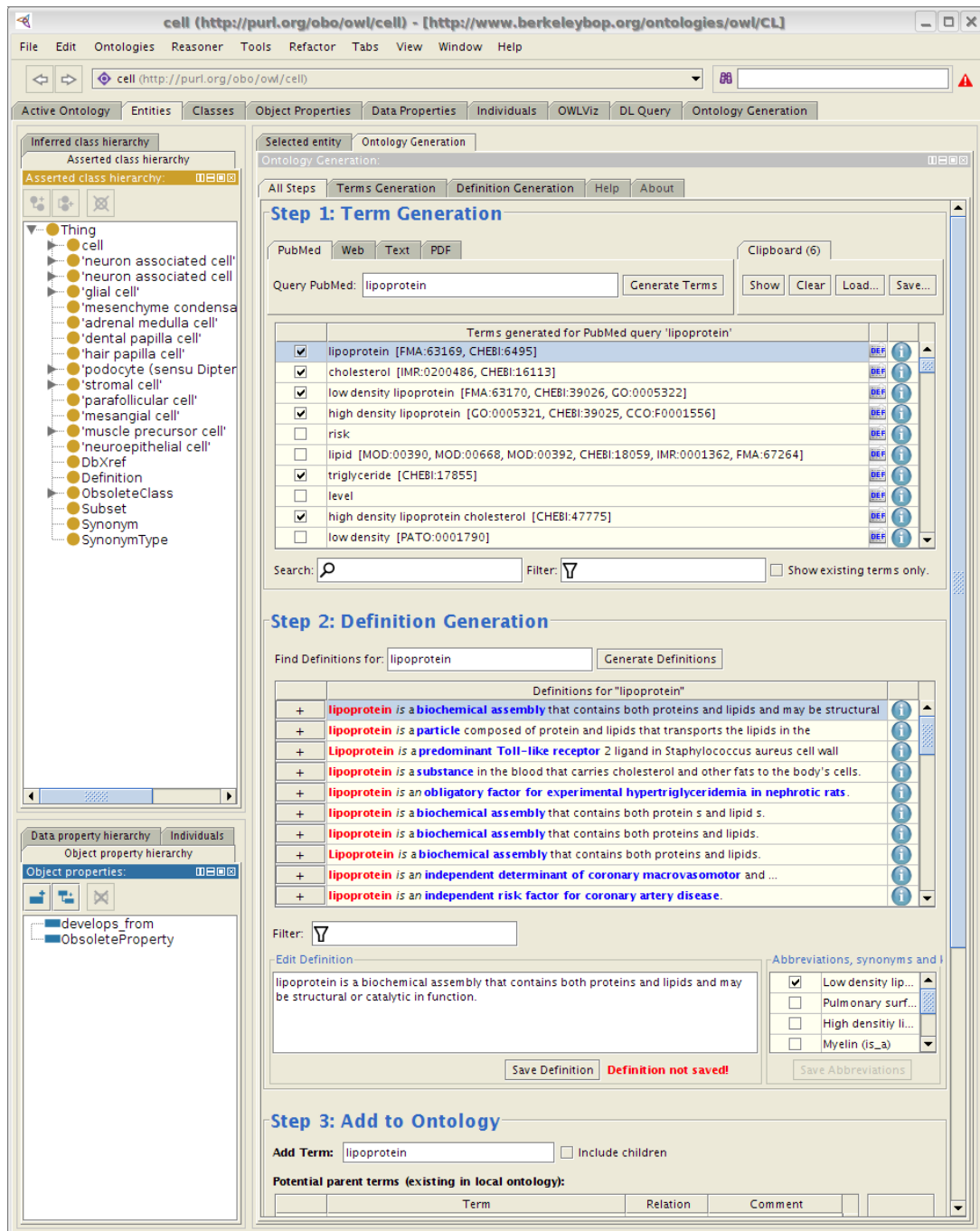


Fig. 7.12. Screenshot of the Ontology Generation Plug-in for Protégé 4.

7.4 Go3R Ontology Editor

OBO-Edit and Protégé are too complicated for many use cases. The Go3R ontology editor has been designed to support the creation of taxonomies for document classification as used in semantic search. The integration of ontology generation methods has been a requirement.

Current ontology editors use knowledge models of varying complexity and differ in scalability and usability. Numerous editors are available to the ontology engineer. There have been five editors reviewed which are used in the life sciences, namely SWOOP³, Protégé⁴, OilEd⁵, pOWL⁶, and OBO-Edit⁷. The editors Protégé, OntoEdit, and OilEd are stand-alone applications for managing ontologies and were compared in Stojanovic and Motik (2002). The editor pOWL is a web-based editor and development platform for the semantic web that supports RDFS/OWL ontologies of arbitrary size. A recent development is Collaborative Protégé, an extension of the existing Protégé system that supports collaborative ontology editing as well as annotation of both ontology components and ontology changes.

Reviewed ontology editors

Protégé is a free, open-source ontology editor for frame-based and OWL ontologies. With **WebProtege** there exists a web-based ontology editor supporting collaborative ontology editing (available under <http://protege.stanford.edu/>).

OilEd (Bechhofer et al., 2001) is an ontology editor for frame-based ontologies and ontologies specified in the web ontology language (OIL). Even though OWL has become a standard in the meanwhile, we reviewed the user interface and the editors functionality has been included in the review.

pOWL is being described as semantic web development platform which is able to manipulate RDFS/OWL ontologies (available under <http://sourceforge.net/projects/powl/>).

OBO-Edit is an open source ontology editor developed and maintained by the Gene Ontology Consortium (Day-Richter et al., 2007). OBO-Edit is optimized for the OBO biological ontology file format. It features an easy to use editing interface, a simple but fast reasoner, and powerful search capabilities (description from <http://oboedit.org/>).

SWOOP is a tool for creating, editing, and debugging OWL ontologies. It was produced by the MIND lab at University of Maryland, College Park and is now an open source project (available under <http://code.google.com/p/swoop/>).

³ <http://code.google.com/p/swoop/>

⁴ <http://protege.stanford.edu>

⁵ <http://oiled.man.ac.uk/>

⁶ <http://powl.sourceforge.net/>

⁷ <http://oboedit.org>.

Editor	SWOOP	Protégé (WebProtégé)	OilEd	pOWL	OBO-Edit
Version	2.3 beta4	3.1.1	3.5.7	0.93	2.1
Functionalities					
add/edit/replace	-/+/-	+/-/-	+/-/+	+/-/-	+/-/-
delete/cut/recursive delete	+/-/-	-/+/-	+/-/-	+/-/-	+/-/-
move/copy/clone	-/+/-	+/-/-	+/-/-	-/-/-	+/-/+
collaborative editing	-	+	-	-	-*
Ontology representation					
list/tree/graph	+/-/+	+/-/-	+/(+)/-	o/+/-	+/-/+
highlight changes	-	-	-	-	+**
Search					
exact, fussy, regex	+/-/-	+/-/+	+/o/-	+/o/-	+/-/+
Versioning					
undo	-	+	-	+	+
redo	-	+	-	-	+

Table 7.2. Overview over the functionalities of selected ontology editors used in the life sciences
 * the simplicity of the OBO format allows external versioning and merging; ** filters and custom renderers available

Functionality

Table 7.2 presents an overview of the functionalities of five existing ontology editors used in the life sciences. Most editors include standard editing capabilities for concepts, properties, instances as well as standard relations, namely concept inheritance based on *part-of* and *is-a* relationships. For all these entities the operations add, remove and modify are supported. All listed operations can be regarded as simple changes (Stojanovic and Motik, 2002) and are implemented in nearly every editor. On the other side there exist composite changes. An example is moving a concept from one parent to another parent. Here the concept's subclass relationship gets deleted and a new subclass relationship to the new parent concept gets created.. It is not possible to replace the task of moving concepts by a sequence of deletions and additions, because the identity of the subject of change itself gets changed.

Beside the described actions, certain constraints have to be maintained by the editor itself. All changes performed within one encapsulated action need to transform the ontology from one consistent and valid state to another one. Accordingly, the constraints consistency and validity need to be ensured. The constraints which have to be met in particular depend strongly on the intended application. The language specification, e.g. for OWL (Patel-Schneider and Horrocks, 2004), is one possible source for consistency constraints. If an ontology is consistent with its language specification it is regarded as *well-formed*. Thus, various notions of consistency can be distinguished (Haase and Stojanovic, 2005).

Definition 7.1 (Structural Consistency). *To ensure structural consistency an ontology needs to obey the constraints of the ontology language.*

Definition 7.2 (Logical Consistency). *An ontology is logically consistent if it is satisfiable, thus does not contain contradicting information.*

Overview over ontology editors see Table 7.2

Definition 7.3 (User-defined Consistency). *Application requirements lead to consistency constraints which have to be met in order to call an ontology consistent.*

All editors must ensure structural consistency. Some systems like OBO-Edit or Protégé address the need for logical consistency and combine editing with consistency checking and validation. Validity ranges from syntactical validity (“Are all concepts and relations declared when used?”) to semantic validity (“Are all concepts or relations used as declared for the ontology?”).

From all reviewed editors only Protégé and OBO-Edit are actively developed. Both are stand-alone software that need to be installed. For the collaborative development of ontologies for search application, one cannot require from each person willing to contribute, to install a special editor. Additionally the synchronisation efforts are too high. The recently released WebProtégé allows collaborative editing, is a good platform, but possibly still too complicated for the averaged user. WebProtégé was only released once the newly created Go3R Editor was already in use. As result of the review of older and existing ontology editors number have been collected for an editor to create the ontology of Go3R (Chapter 8) and other ontology-based search engines. An ideal editor should be usable without additional training and provide editing capabilities commonly used, e.g. in file managers. Instead of adding a relation, an object should be created below or with reference to some other term/class. Instead of deleting a relation or adding a new relation, the change should be performed with reference to the term/class as copy, cut, paste, or drag&drop operation. None of the existing editors provides this simple user interaction in combination with collaborative editing.

Requirements

1. **Collaborative ontology editing:** Several users should be able to manipulate the ontology at the same time. The editor should allow community-driven ontology development without additional efforts for synchronising different versions and releases.
2. **Web-based interface:** The editor should be integrated in the Go3R search engine. A web-based user interface would allow integration in the search engine and lower the barrier of users to contribute to the development.
3. **Keyboard control:** The editor should support full keyboard control to comply to the customs of the intended users. Operations, such as create, delete, copy, paste, and edit term should be supported.
4. **Immediate change propagation:** Changing the ontology should have direct impact on the document retrieval in the search engine. This enables the ontology developer to validate the classification performance, starting with the next search.
5. **Primary support for subclass relationships:** The classification (navigation) structure in Go3R is a tree, where subsumption relationships, like *part_of* and *is_a* are treated as subclass relationships. The primary goal is to create the taxonomic backbone of the ontology which is required for ontology-based search applications.
6. **Extensibility to distinguish between different types of relationships:** Some applications require a richer ontology model. Especially the task of locating con-

cepts in natural language text, can benefit from a richer ontology model for sense disambiguation and topic classification.

7. **Multiple ontologies:** For text retrieval, several independent ontologies need to be used and edited at the same time. Go3R uses the Go3R Ontology and an Nanotoxicology Ontology developed for the *Federal Institute for Risk Assessment (BfR) Berlin, Germany*, both need to be technically and visually distinguished.
8. **Incorporation of ontology generation:** A requirement for the ontology editor GUI is the seamless integration of predicted ontology terms.
9. **Ontology generation from different text sources:** Ontology generation from text relies on text sources. These text sources need to be selectable in the ontology editor.
10. **Ontology generation from different ontologies:** Ontology generation from text relies on existing ontologies. These ontologies need to be selectable in the ontology editor.

Functional and representational requirements listed above led to the first design study of the Yggdrasil Ontology Editor shown in Figure 7.13.

7.4.1 Design study for a new web-based ontology editor

The first step towards the development of the Go3R Ontology Editor has been a design study to clarify, how to integrate ontology generation (**Requirement 8**) with taxonomy editing (**Requirement 2**) in a simple and intuitive user interface. The hierarchical structure of the ontology is represented by columns, where the leftmost column has a special status. This left column contains sliding bars, search functionality, and status information. The top sliding bar holds selectable items representing ontologies (**Requirement 10**) and document sources (**Requirement 9**). By selecting an ontology or document source this is employed as textual or semantic source for the prediction of novel terminology. Each column to the right has a menu bar to access the change operations “New”, “Edit”, and “Delete” for the selected term. The red menu point “Predict” has been added to initiate the integrated term prediction (**Requirement 8**). The prediction method in this study has been intended to generate terms in the respective column using the parent terms as well as the existing siblings. The icons preceding a term qualify whether a term was extracted from text or has been imported from other ontologies. The design study addresses all functional requirement and allows the implementation of the technical requirements, such as keyboard control and change propagation.

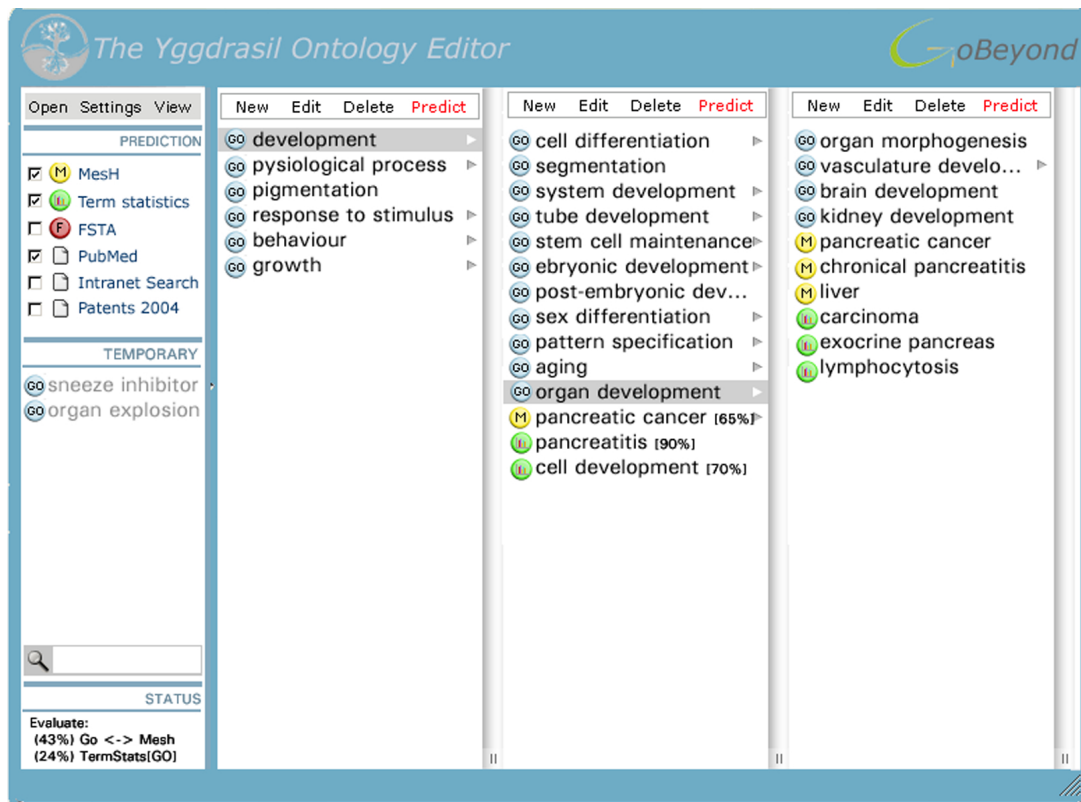


Fig. 7.13. Design study of the Yggdrasil Ontology Editor. The graphical user interface is designed to select the ontologies to be viewed, edited, and referenced, and to select the document sets used as source for ontology generation. The hierarchy is shown in columns, that at each column generated terms can be automatically added. A clipboard, here TEMPORARY, holds copied and archived terms. All basic commands (new, edit, delete, predict) can be directly accessed.

7.4.2 The Ontology Editor (Version 1)

In this first implementation of a web-based editor (**Requirement 2**) the support multiple relationships (**Requirement 6**) has been added by extending the column model to contain several sections per column. For simple taxonomies (**Requirement 5**) with only one relationship a column is separated in an upper part, showing the parents, and a lower part showing the children of the selected term in the column to the left. Additional relationships will be supported by either separating a column in several parts or by switching between relationship types. Multiple ontologies can be loaded (**Requirement 7**). The ontologies are displayed in the left most column in the top slider. The editor is integrated in Go3R, which means that the ontology of the search engine can be directly edited (**Requirement 4**). This integration in the web application ensures, that only one single copy of the ontology exists and together with the “transaction save ontology model” simultaneous editing is made possible (**Requirement 1**). The “Edit Term Dialog” is reduced to only the required input fields displaying the known id of a term and its parents and allowing the user to specify a label, description, synonyms, see Figure 7.14. The ontology editor can be controlled entirely with keyboard, which increases productivity (**Requirement 3**). The editor allows composite operations like move (delete relation + add relation) by drag&drop and copy (create term + add relation). This first implementation supports all requirements apart from the integration of ontology generation (**Requirements 8, 9, and 10**).

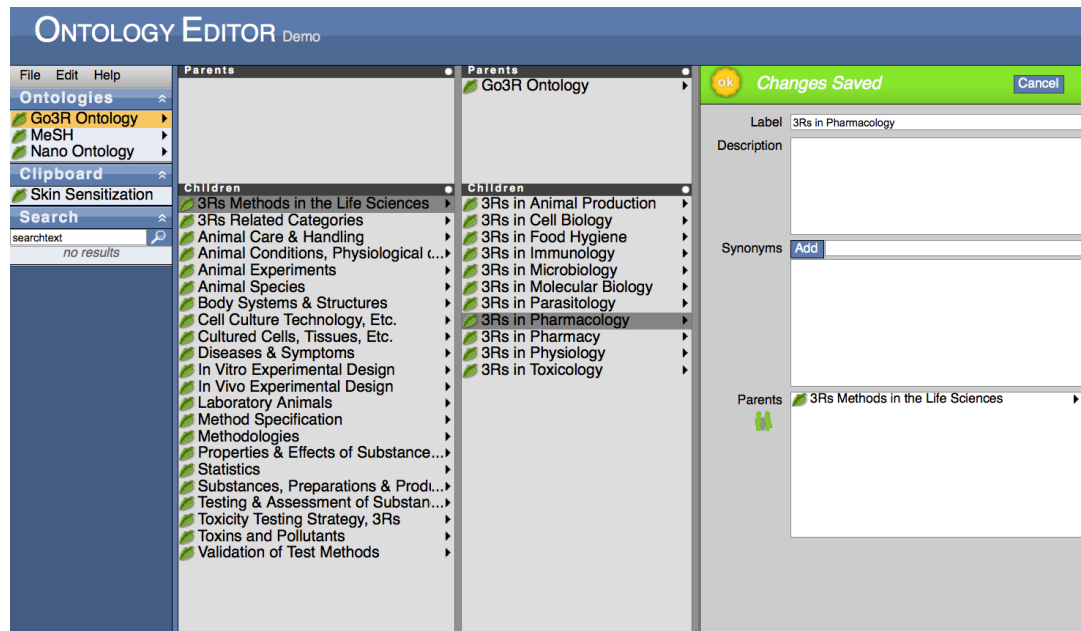


Fig. 7.14. Screenshot of the web-based Ontology Editor (Version 1). Other than the design study, each column is splitted to show multiple parents of a term. The ontology shown in this example is the initial version of the Go3R ontology. The editor has been implemented with the help of HicknHack Software – Andreas Reischuck | Maik Lathan | Michael Starke GbR and is available under <http://www.hicknhack-software.com/software/ontologyeditor/>

7.4.3 The Go3R Ontology Editor (Version 2)

The Go3R Ontology Editor in its current state is a re-implementation in a different web framework and follows the same specification as version 1. The Go3R Ontology Editor supports a number of change operations to make the editor more user friendly. The supported operations are listed in (Table 7.3) and are displayed in the context menu in Figure 7.17. Unique to the Go3R editor in terms of functionality is the “recursive delete” function not present in any of the other editors (Table 7.4). The term generation has been integrated as separate view (Figure 7.16).

New term	creating a new term and a relation to the new parent at the location the term was created
Edit term	opening the edit dialog to modify label, description, or synonyms of a term
Insert Term List	insert several terms in one action
Copy	copying there term and places it in the clipboard
Paste	creating a new parent relationship from the pasted term to the location where it has been pasted
Remove from parent	deleting the parent-child relationship and places the term in the clipboard if no other parent exists
Delete	deleting the term; if a term is removed from its last parent term, the term gets destroyed
Delete recursive	deleting all terms subsuming the term by removing the relations. Terms which do not have any parent after this operation are being destroyed.

Table 7.3. Supported change operations of the Go3R Ontology Editor as shown in Figure 7.17.

Editor Version	SWOOP 2.3 beta4	Protégé 3.1.1	OilEd 3.5.7	pOWL 0.93	OBO-Edit 2.1	Go3R-Editor 2.0
Functionalities						
add/edit/replace	-/+/-	+ /+/-	+ /+ /+	+ /+/-	+ /+/-	+ /+ /+
delete/cut/recursive delete	+ /- /-	- /+ /-	+ /+ /-	+ /- /-	+ /+ /-	+ /+ /+
move/copy/clone	- /+ /-	+ /+ /-	+ /+ /-	- /- /-	+ /+ /+	+ /+ /-
collaborative editing	-	+ ¹	-	-	- ²	+
Ontology representation						
list/tree/graph	+ /- /+	+ /+ /-	+ /(+) /-	o /+ /-	+ /+ /+	- /+ /-
highlight changes	-	-	-	-	+ ³	-
Search						
exact, fussy, regex	+ /+ /-	+ /+ /+	+ /o /-	+ /o /-	+ /+ /+	+ /+ /-
Versioning						
undo	-	+	-	+	+	- ⁴
redo	-	+	-	-	+	-

Table 7.4. Overview over the functionalities of selected ontology editors used in the life sciences; ¹ only WebProtégé; ² the simplicity of the OBO format allows external versioning and merging; ³ filters and custom renderers available; ⁴ versioned ontology model, not accessible from editor

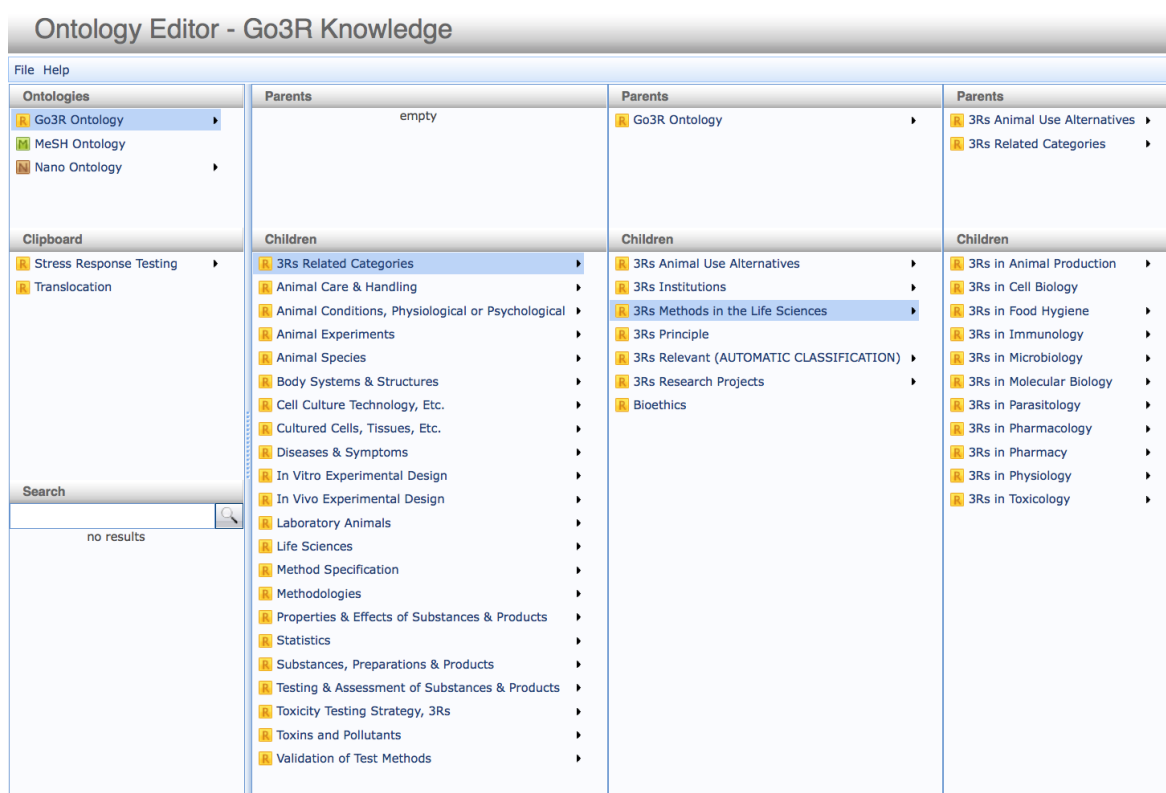


Fig. 7.15. Web-based Go3R Ontology Editor (Version 2). The editor has been integrated in Go3R and allows to change the ontology in the online version of the search engine. Changes in the ontology lead to re-annotations of affected documents and immediately affects subsequent searches.

Term Generation			
skin sensitization			
<input type="button" value="Generate From Text"/> <input type="button" value="Generate From PubMed Query"/>			
<input type="checkbox"/> sensitization	4.881		sensitizations,sensitization,Sensitization
<input type="checkbox"/> skin	7.906		skin,Skin
<input type="checkbox"/> allergen	10.267		allergen,Allergens,allergens
<input type="checkbox"/> allergy	10.499		Allergy,allergy
<input type="checkbox"/> test	11.598	TST,test	Tests,testing,tests,test
<input type="checkbox"/> skin prick test	15.099	SPT,SPTs	skin prick testing,skin-prick tests,Skin prick testing,Skin prick test
<input type="checkbox"/> asthma	18.034		Asthma,asthma
<input type="checkbox"/> pollen	24.342		Pollens,pollen,pollens
<input type="checkbox"/> exposure	25.960		exposure,exposures,Exposure
<input type="checkbox"/> atopic	27.569		atopics,atopic,Atopic
<input type="checkbox"/> children	27.836		Children,children
<input type="checkbox"/> food	28.055		Food,food,foods
<input type="checkbox"/> contact	30.574		contacts,contact
<input type="checkbox"/> hypersensitivity	31.579		hypersensitivity,Hypersensitivity
<input type="checkbox"/> dermatitis	32.567		dermatitis
<input type="checkbox"/> atopic dermatitis	32.973	AD	atopic dermatitis,Atopic dermatitis,Atopic Dermatitis
one row			
<input type="button" value="Create Terms"/> <input type="button" value="Close"/>			

Fig. 7.16. Term generation within the Go3R Ontology Editor. Terms have been generated from text retrieved from PubMed for the query skin sensitization. The table shows in each row the term label, a score, abbreviations (if applicable), and lexical variants found in the texts.

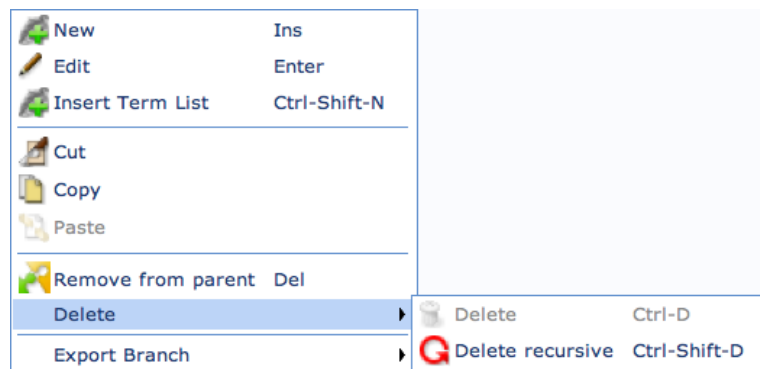


Fig. 7.17. Term context menu of the Go3R Ontology Editor. The menu shows the supported operations “New concept”, “Edit concept”, bulk creation of a several terms (“Insert Term List”). The operations “Cut”, “Copy”, “Paste”, “Remove from parent” help to manipulate the relation to parent terms. “Delete” removes a term. “Delete recursive” removes a whole branch.

ok Changes Saved Save Cancel

Id 827

Label EpiDermTM

Description [GENERATED DEFINITION] EpiDermTM is a commercially available human skin model consisting of normal human-derived epidermal keratinocytes (NHEK), which have been cultured to form a multilayered, highly

Synonyms Add

EpiDerm

EpiDerm(TM)

OECD TG 431

EPI-200

OECD 431

Parents

- Human Skin Model Test
- In Vitro Human Skin Assays
- Organotypic Tissue Cultures

Fig. 7.18. Edit term dialog of the web-based Go3R Ontology Editor. It displays the ID and parents of a term and allows to change label, description/definition, and synonyms.

7.5 Web-based Term Generation Platform

Independently from ontology editors, the web-based term generation platform provides support for the collaborative acquisition of terminologies. Texts, either provided or retrieved from PubMed or the ZEBET database are evaluated and the terminology is getting extracted and displayed in list form. Syntactic grouping has been added to pre-organise the retrieved terminology. A group contains all terms belonging to the closure created over tokens of the terms label or synonyms. A term is added to a group of terms if one of its tokens overlaps with tokens of some other term in the group. The groups are ranked based on the domain relevance of the best (most relevant) representative. Terms within a group are ranked likewise by relevance.

Enter PubMed query, PubMed IDs, Zebet IDs or copy/paste plain text

cholesterol

Submit as Text

Submit as Zebet Query

Submit as PubMed Query

Context for curation (any descriptive phrase)

unilever_colworth

Load All Terms

Show contexts

☒ View positive marked ☒ View negative marked ☒ View undecided

Export displayed terms

ranked by score

cholesterol	[Ch, CHO]	[cholesterol, Cholesterol]	
non-high-density lipoprotein	[HDL]	[non-high-density lipoprotein]	
lipid		[Lipids, lipid, Lipid, lipids]	
metabolic syndrome	[MS, MetS]	[metabolic syndrome, Metabolic Syndrome, metabolic syndromes, METABOLIC SYNDROME, Metabolic syndrome]	
high-density lipoprotein	[HDLs, HDL]	[high-density lipoprotein, High-density lipoprotein, High-density lipoproteins, high-density lipoproteins, high density lipoprotein, high density lipoproteins]	
low-density lipoprotein	[LDLs, LDL]	[Low-density lipoprotein, Low density lipoprotein, low-density lipoproteins, low density lipoprotein, low density lipoproteins, Low Density Lipoprotein, low-density lipoprotein]	
risk		[Risk, risk, risks]	
lipoprotein		[lipoprotein, lipoproteins]	
triglyceride	[TG]	[Triglycerides, triglyceride, triglycerides]	
(hs)-C-reactive protein	[CRP]	[CRP, (hs)-C-reactive protein]	
LDL		[LDL-, LDL]	
metabolic		[metabolic, Metabolic]	
density		[density]	

Fig. 7.19. The web-based Term Generation Platform has been developed for the collaborative acquisition of terminology from text. Each row shows label, abbreviations, and lexical variants found in text.

In a semi-automatic fashion the user can select or reject terms. The manual curations can be stored for a user-defined context. The curations for each context are remembered and generated terms in following sessions are displayed as curated previously in the selected context. The curated terms for a context can be loaded and exported for bulk import in other tools.

The term generation platform builds on the same synchronous and asynchronous web services developed for OBO-Edit and Protégé.

7.6 User Scenario – Biocuration

The stepwise example (Section 7.2.1) demonstrates how the creation and extension of ontologies is supported. Additionally, DOG4DAG can be used directly for biocuration. By searching for a gene product the relevant terms from known ontologies and novel terms are suggested. This helps biocurators directly for the annotation of genes and gene products but also indirectly to identify other relevant ontologies and terms to be included in the ontology.

Example 7.4 (Pax6). Let's consider the gene product Pax6. Pax6 is a "*transcription factor playing a crucial role in the development of the eye*" according to DOG4DAG's definition generation. Querying with Pax6 brings up terms such as "*eye*", "*development*", and "*aniridia*". The first generated definition of the latter states that it is "*a disease in which the iris fails to form normally*". The entry for aniridia also provides references to the disease ontology, GO biological processes and the human phenotype ontology (Robinson et al., 2008). With Gene Ontology loaded in OBO-Edit, all generated terms similar to GO terms are shown in **bold face**. For Pax6 they are "*developmental process*", "*transcription*", "*neurogenesis*", or "*eye development*". They correspond one-to-one to the UniProtKB⁸ annotations of human Pax6, which we use for validation.

Example 7.5 (Eya1). The gene Eya1 is a "*homolog of eyes absent in Drosophila and essential for various organ formations in vertebrates*" according to DOG4DAG'S first generated definition. Querying with Eya1 brings up terms such as "*Branchio-Oto-Renal syndrome*", "*development*", "*ear*", "*eye*", "*kidney*", "*Pax*", and "*Hox*". The tool offers "*BOR*" as abbreviation for the syndrome and a click on the information icon retrieves papers linking Eya1 and the syndrome to kidney and ear development such as "*EYA1 Mutations associated with Branchio-Oto-Renal Syndrome Result In Defective Otic Development in Xenopus laevis*" (PMID 19951260). The link of Eya1 to the Hox and Pax genes is also found by clicking the information icon on either Hox or Pax bringing up the paper "*A Hox-Eya-Pax complex regulates early kidney developmental gene expression.*" (PMID 17785448).

Further, the usefulness of the Ontology Generation Tool will be presented on examples of recent ontology request submitted to the Gene Ontology project tracker. Each request will be briefly summarised and the support available through DOG4DAG will be highlighted.

⁸ www.uniprot.org

Example 7.6 (Annotation issue — genes associated with GO:0006488 (LLO synthesis) - ID: 2978670; 2010-03-29).

The reporter of the annotation request submitted a list of genes to be annotated with “*dolichol-linked oligosaccharide biosynthetic process*” (GO:0006488).

Suggested annotations:

- all genes: DPAGT1, ALG1, ALG11, RFT1, ALG3, ALG9, ALG10A/B, ALG5, DPM1, DPM2, DPM3, PMM1, PMM2, GMPPA, GMPPB, PGM3, UAP1, DOLK1
- genes already annotated in GO: ALG12, ALG2, DPAGT1, MPDU1

We checked for all these genes, whether “*dolichol-linked oligosaccharide biosynthetic process*” is contained in the list of generated terms and at which position. For 11/21 genes the suggested GO term was recognised (Table 7.5). Recognising the GO term required in this case that first the complex noun phrase “*dolichol-linked oligosaccharide biosynthetic process*” was extracted correctly. Second, the mapping to GO was performed correctly, and third the term was ranked high. Further we tested for the four already annotated genes ALG12, ALG2, DPAGT1, MPDU1 whether any of the other manually curated GO annotations in UniProtKB can be discovered with DOG4DAG (Table 7.5).

	Gene name	found GO:0006488	Position
Comparison with manual GO annotations in UniProtKB for genes ALG12, ALG2, DPAGT1, MPDU1	ALG12	✓	9
	ALG2	✓	9
	DPAGT1	✓	5
	MPDU1	✓	4
ALG 12: 2 of 3 annotations found <i>dolichol-linked oligosaccharide biosynthetic process;</i> <i>protein folding</i>	ALG1	✓	13
	ALG11		
ALG2: 2 of 3 annotations found <i>dolichol-linked oligosaccharide biosynthetic process;</i> <i>protein amino acid glycosylation in endoplasmic reticulum ≈ glycosylation + endoplasmic reticulum</i>	RFT1	✓	9
	ALG3	✓	11
	ALG9	✓	6
	ALG10A/B		
	ALG5		
DPAGT1: 0 of 2 annotations found <i>UDP-N-acetylglucosamine-dolichyl-phosphate</i> N- <i>acetylglucosaminophosphotransferase activity ≈ mUDP-GlcNAc:Dolichol Phosphate N-Acetylglucosamine-1 Phosphate Transferase</i>	DPM1	✓	15
	DPM2	✓	7
	DPM3	✓	3
	PMM1		
MPDU1: 1 of 1 annotation found <i>dolichol-linked oligosaccharide biosynthetic process</i>	PMM2		
	GMPPA	no documents	
	GMPPB	no documents	
	PGM3		
	UAP1		
	DPOLK1	no documents	

Table 7.5. Listing for genes suggested in a Gene Ontology Annotation issues tracker requested for annotation with GO term “*dolichol-linked oligosaccharide biosynthetic process*” (GO:0006488)

Example 7.7 (Ontology request — Luteolysis - definition, parentage and children - ID: 3001076; 2010-05-13).

Request for a new definition for luteolysis:

luteolysis – The lysis or structural demise of the corpus luteum. During normal luteolysis, two closely related events occur. First, there is loss of the capacity to synthesize and secrete progesterone (functional luteolysis) followed by loss of the cells that comprise the corpus luteum (structural luteolysis)

Request for new children “functional luteolysis” and “structural luteolysis” with definitions:

functional luteolysis – in which the corpus luteum loses the ability to produce progesterone to allow development of new follicles

structural luteolysis – removal of the inactive corpus luteum involving apoptosis, occurs after the start of functional luteolysis

A query luteolysis is submitted to PubMed. The generated terms contain the proposed children “structural luteolysis” and “functional luteolysis” and other potential child terms such as “induced luteolysis” and “spontaneous luteolysis” after filtering with luteolysis. The generated terms are listed in Table 7.6. The definition extraction finds correct definitions for “luteolysis”, “structural luteolysis”, and “functional luteolysis” which are similar to the definitions suggested by the user in the submitted ontology request. The generated definitions are listed in Table 7.7.

Rank	Generated term
1	functional luteolysis
2	induced luteolysis
3	structural luteolysis
4	spontaneous luteolysis
5	PGF(2alpha)-induced luteolysis

Table 7.6. Term candidates for PubMed query luteolysis. The 1st and 3rd terms are the proposed children in the GO Ontology Request ID: 3001076

Term	Generated definition
luteolysis	Luteolysis is the structural and functional degradation of the corpus luteum (CL) that occurs at the end of the luteal phase of both the estrous and menstrual cycles in the absence of pregnancy.
functional luteolysis	Functional luteolysis refers to suppression of progesterone synthesis and secretion, and precedes structural luteolysis.
structural luteolysis	Structural luteolysis is a complex process responsible for the elimination of the corpus luteum (CL).

Table 7.7. Extracted definitions for terms luteolysis, structural luteolysis, and functional luteolysis. All three definitions show significant overlap with the proposed definitions in the GO Ontology Request ID: 3001076.

Example 7.8 (Ontology request — NTR chloride-activated potassium channel activity - ID: 2204957; 2008-10-28).

Request for a new term:

The reporting user asks to add “chloride-activated potassium channel activity” as child of GO:0005267 (potassium channel activity).

In the GO there exist 15 terms with a label or synonym ending with “potassium channel activity” (Table 7.8). With a PubMed search for potassium channel activity 12 terms ending with potassium channel activity have been generated. Five are already GO terms (Table 7.9). “M-current potassium channel activity” is a synonym for “voltage dependent potassium channel activity”. Five terms are good candidate terms.

ID	GO term label
GO:0005250	A-type (transient outward) potassium channel activity
GO:0015272	ATP-activated inward rectifier potassium channel activity
GO:0015269	calcium-activated potassium channel activity
GO:0070089	chloride-activated potassium channel activity
GO:0005251	delayed rectifier potassium channel activity
GO:0015467	G-protein activated inward rectifier potassium channel activity
GO:0022894	Intermediate conductance calcium-activated potassium channel activity
GO:0005228	intracellular sodium activated potassium channel activity
GO:0005242	inward rectifier potassium channel activity
GO:0060072	large conductance calcium-activated potassium channel activity
GO:0005252	open rectifier potassium channel activity
GO:0015271	outward rectifier potassium channel activity
GO:0005267	potassium channel activity
GO:0016286	small conductance calcium-activated potassium channel activity
GO:0005249	voltage-gated potassium channel activity

Table 7.8. Listing existing terms in the Gene Ontology that end with “potassium channel activity”.

ID (mapped):	Generated term label
GO:0005267	potassium channel activity <i>ATP-sensitive potassium channel activity</i> <i>single potassium channel activity</i>
GO:0015269	calcium-activated potassium channel activity <i>adenosine triphosphate-sensitive potassium channel activity</i> <i>M-current potassium channel activity</i>
GO:0005251	delayed rectifier potassium channel activity <i>modulation of potassium channel activity</i>
GO:0060072	large-conductance calcium-activated potassium channel activity <i>astrocyte potassium channel activity</i> <i>MthK potassium channel activity</i>
GO:0005249	voltage-gated potassium channel activity

Table 7.9. Listing of 12 terms generated with the OBO-Edit Ontology Generation Tool for the query “potassium channel activity”. Five terms ending with “potassium channel activity” exist in the Gene Ontology (**bold**), another six are good candidate terms (*italic*).

7.7 Summary and Discussion

The DOG4DAG ontology generation methods developed in this thesis have been seamlessly integrated in OBO-Edit and Protégé, two widely used ontology editors in the life sciences. The systems offers to either submit the query to PubMed or the web or to upload text or PDF documents. While PubMed is the default source for terminology, the web is often useful since full-text articles can be implicitly included in the search. State-of-the-art natural language processing techniques are used to rank the relevance of terms for the domain to be modeled. Results in Chapter 3 (Terminology Generation) showed that term generation can improve the completeness of an ontology by suggesting up to 75% good candidate terms in the top 50 ranked terms.

When adding a term to the ontology, possible parents are suggested on the basis of generated definitions and existing terms from other ontologies are cross-referenced. Public resources such as Wikipedia, full text articles and web sites are incorporated to generate definitions, which follow the well defined structure *A is a B with property C* if available. The results in Chapter 4 (Definition Extraction) reveal that for 78% (MeSH) and 55%(GO) a suitable definition could be suggested within the top 10 ranked definition. For this evaluation 1,000 terms from GO and MeSH have been randomly selected. Based on the generation of good definitions it is possible to suggest the likely parent of *A*, namely *B*. The results in Chapter 5 (Taxonomy Generation) showed that for 54% of terms in MeSH and 38% of terms in GO, correct relations to ancestors could be predicted.

In the following we will discuss DOG4DAG in relation to other tools supporting aspects of automatic ontology creation within ontology editors and will take position on how DOG4DAG's input complies to the design guidelines proposed by Schober et al. (2009).

7.7.1 Ontology learning tools

Over the past few years some text-mining approaches and systems for ontology learning have been developed such as TerMine, Text2Onto, OntoLT for Protégé, or OntoLearn.

TerMine based on the C-Value method (Frantzi et al., 2000) retrieves and ranks multi-word phrases. Since 15% of all MeSH terms and synonyms, as well as most gene names are consisting of a single word, DOG4DAG's inclusion of single words as terms is an important extension not present in TerMine. DOG4DAG achieves this by ranking terms according to their relative importance. The grouping of all lexical variants and abbreviations leads to significant frequency counts. Text2Onto (Cimiano and Völker, 2005) is an ontology learning framework including a graphical user interface which supports terminology recognition, hypernymic and mereological relationship extraction. The OntoLT Protégé Plug-in⁹ includes rule based extraction of candidate terms and relations based on linguistic features of provided texts. Both systems build on strong linguistic foundations but require user input prior to the generation of terms or relations, such as the creation of rules in Text2Onto and an

⁹ olp.dfki.de/OntoLT/OntoLT.htm

annotated corpus of documents for OntoLT (Buitelaar et al., 2004). Our evaluation in Alexopoulou et al. (2008) showed, that the term generation of DOG4DAG performs equally or better than the state-of-the-art systems Text2Onto and TerMine.

In comparison to all other systems, DOG4DAG uniquely combines the generation of terms and taxonomic relations with definition extraction and simple ontology mapping in a ready-to-use tool for the life science community.

7.7.2 Design guidelines

An important question is whether generated terms satisfy naming guidelines proposed for manually created terms as put forward by Schober et al. (2009). The authors comprehensively evaluated existing open biomedical ontologies and defined a number of guidelines for naming concepts to reach acceptance by the community. This is satisfied by all term generation approaches since they are based on text, which should be the output of the community in question. In DOG4DAG this is additionally satisfied by supporting generation of terms from PubMed abstracts, web queries, text files, and repositories of PDF documents. According to Schober et al., abbreviations should be captured with the terms. This is indeed the case in DOG4DAG, which groups variations of terms and their abbreviations. Schober et al. promote the avoidance of ambiguity. In DOG4DAG, ambiguous terms are easily identified through their generated definitions and can hence be avoided. Schober et al. also recommend to avoid negations and conjunction in terms. Since these are rarely used directly in text, DOG4DAG does not suffer from this problem. E.g. for in total 420,000 terms generated for the evaluation in Section 3.4.1 (Noun phrases as term candidates) only 10 contained the words *without*, *excluding*, or *not* (negation) and only 462 the word *and* (conjunction). Schober et al. also emphasise the importance of term re-use. DOG4DAG supports this by grouping variants of terms and checking whether they exist in existing OBO ontologies, in which case the label of the OBO term is recommended and it is offered to include its descendants, too.

7.7.3 Biocuration

It has been shown on recent examples in Section 7.6 (User Scenario – Biocuration) that the OBO-Edit Ontology Generation Tool can support the Gene Ontology annotation and the associated extension of the Gene Ontology. For instance,

- 11 of 21 annotations for “*dolichol-linked oligosaccharide biosynthetic process*” could be automatically found in abstracts;
- the children of luteolysis were correctly found and luteolysis and the children were semi-automatically defined;
- five new and five known types of potassium channel activity were found with the term generation.

In Winnenbourg et al. (2008) we concluded that manual curation of literature is necessary for high-quality annotation but can be supported by automated methods. Systems like Textpresso (Müller et al., 2004) successfully support manual curation and recently have been estimated to speed up the curation process of *C.elegans* proteins to GO cellular components by at least 8-fold (Van Auken et al., 2009).

Integrated in the GO annotation process described by Hill et al. (2008), DOG4DAG helps to identify appropriate ontology annotation terms, by showing the GO terms used in literature and in the same way collecting the literature reference to include in the annotation record. In cases where novel terms need to be created DOG4DAG will help to define and place the new term in the GO. A distinctive feature in contrast to other systems is that the subject of annotation can be changed by simply loading a different ontology into the ontology editor.

Definitions of terms in ontologies are important, but cumbersome to define. As Table 4.6 showed nearly all GO and MeSH are defined. However, for more specialised ontologies, this is not the case. In the over 90 OBO ontologies there are 99,418 terms without a definition. Thus, there is a huge potential to save manual labor when defining terms with DOG4DAG.

7.7.4 Taxonomy editor for Go3R

In addition, a novel collaborative taxonomy editor has been specified as user-friendly, web-based alternative to existing ontology editors. It allows domain experts to contribute to the Go3R ontology (Chapter 8) without having to install or learn new complex software systems. The editor directly modifies the ontology of the semantic search engine which immediately has effect on subsequent searches.

7.7.5 Limitations

There are five major limitations: The reproducibility of web derived results, the ability to compose terms and to extract specific relations, the completeness of definitional sentences, and the incomplete mapping to existing ontology terms.

Reproducibility of results: All steps relying on the web documents are not necessarily reproducibility as the index of the web search engine can change. Caching could be a solution, but has not been addressed so far.

Composition of terms: Currently, there are many efforts to understand the composition of ontology terms following patterns (Ogren et al., 2004; Mungall, 2004). In Aranguren et al. (2008) the authors discussed two design patterns for terms. DOG4DAG does not support such a composition process. However, DOG4DAG's filtering of terms helps to realize the value partition pattern. For example, after a search for "stem cell" one can filter to keep only terms containing "stem cell" obtaining among others the value partition mesenchymal, hematopoietic, and neural.

Extraction of specific relations: The second limitation of DOG4DAG is the extraction of relations as promoted in (Smith et al., 2005a; Soldatova and King, 2005). The latter, also mentions that ontologies should contain axioms. DOG4DAG only deals with extraction of parent-child relationships. Since part of speech tagging is used it is in principle possible to extract relations from verb phrases. But since this requires both terms to appear in one sentence, the coverage would be much lower and is therefore currently omitted.

Completeness of definitions: Snippets obtained from web search engines are usually truncated within sentences. Often only partial definitional sentences can be retrieved. Nevertheless, once a good partial definition is found and annotated in a snippet, the full definition is extracted from the original web site or web document if available.

Mapping to existing ontology terms: Currently two technically independent mechanisms are employed to create the mapping from generated terms to existing terms. For referencing to existing terms an external service from the European Bioinformatics Institute is used. We do not have control or knowledge over the versions of the provided ontologies. The mapping to the locally loaded ontology in the editor is performed by the plug-in itself and is optimised for the Gene Ontology. It can happen that mappings are only found by one of the methods.

7.8 Future Extensions

The development of the tools presented in this chapter is not finished. A number of future extensions have been identified and will be included in due time.

Ontology generation integration in OBO-Edit and Protégé For the ontology generation support within the ontology editors future versions will use the context provided by the ontology to rank terms with additional domain relevance. A single ontology model with confidence-weighted relations will incorporate generated relations as well as all existing terms and relations provided by the EBI Lookup Service, the BioPortal services. The usage of the OBO-Edit reasoner will enable identify predictions which implicitly exist or are not allowed.

Go3R Ontology Editor The web-based editor will be extended to include automatic suggestions of new siblings and relations as proposed in the design study (Section 7.4.1). Also definition extraction and ontology look-up are still open for implementation.

7.9 Contributions

OBO-Edit Ontology Generation Tool *Design and implementation, including the underlying web services and the integration within OBO-Edit. The GUI implementation has been supported by the students Atif Iqbal, Götz Fabian, and Marcel Hanke.*

Protégé Ontology Generation Plug-in *Based on the OBO-Edit Ontology Generation Tool the Protégé Plug-in has been developed with the help of Götz Fabian, and Marcel Hanke.*

Go3R Ontology Editor *Idea and specification of the user interaction and GUI layout for an easy to use editor for simple ontologies to create the background knowledge for Go3R. The first editor software has been implemented by Hick'n'Hack Software¹⁰, the second version of the editor has been implemented by Matthias Zschunke (Transinsight GmbH).*

Idavoll Term Generation Platform *Design and implementation of the web site using the GWT Toolkit from Google Inc.*

¹⁰ <http://www.hicknhack-software.com>

Go3R – Semantic Search for Alternative Methods to Animal Testing

References

- Sauer, U. G.*, Wächter, T.*, Grune, B., Doms, A., Alvers, M. R., Spielmann, H., and Schroeder, M. (2009). Go3R - semantic internet search engine for alternative methods to animal testing. *ALTEX*, 26(1):17–31, *Impact factor 2009: 1.3* *shared first author)
- Alexopoulou, D.*, Wächter, T.*, Pickersgill, L., Eyre, C., and Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*, 9(Suppl 9):S2, *Impact factor 2009: 3.7* *shared first author
- Wächter, T., Alexopoulou, D., Dietze, H., Hakenberg, J., and Schroeder, M. (2007). *Anatomy Ontologies for Bioinformatics, Principles and Practice*, volume 6, chapter Searching Biomedical Literature with Anatomy Ontologies, pages 177–194. Springer Computational Biology.
- Dietze, H., Alexopoulou, D., Alvers, M. R., Barrio-Alvers, B., Doms, A., Hakenberg, J., Mönnich, J., Plake, C., Reischuk, A., Royer, L., Wächter, T., Zschunke, M., and Schroeder, M. (2008). *GoPubMed: Exploring PubMed with Ontological Background Knowledge*, chapter Bioinformatics for Systems Biology, pages 385–399, Humana Press.

The goal was to design the new search engine Go3R that is able to retrieve literature describing reduction, replacement, or refinement methods for animal testing according to the 3Rs principle (Russell and Burch, 1959) and distinguish relevant literature from other general biomedical literature. The work included the collaborative creation of the Go3R ontology, and has been shown on examples that the ontology generation methods developed in this thesis are applicable in the domain of animal testing alternatives. They can support the ontology engineer with relevant terms, relevant abbreviations, lexical variants, and definitions. The definition generation method can find definitions for half of the 152 alternative methods for toxicity testing which exist in Go3R.

Go3R is built with the experience gained during the development of a number of earlier systems in different domains which will be described in addition to Go3R in this chapter.

Go3R is an ontology-based search engine — the worldwide first internet search engine for alternative methods to animal experiments — which directly addresses the need for information retrieval that arises from

- a) the regulations of EU Directive 86/609/EEC for the protection of laboratory animals, which obliges scientist to consider whether a planned animal experiment can be replaced, reduced, or refined.

- b) *the new EU Chemicals Regulation REACH which is expected to lead to an EU-wide increase in the numbers of animals used in animal experiments of up to 400,000 animals per year.*
- c) *the diversity of relevant information contained in estimated 50 Million potentially relevant documents which are scattered across the internet, patent databases, literature databases and intranets.*

The Go3R project is a collaboration between the Bioinformatics Group of the Technical University of Dresden, Germany; Scientific Consultancy - Animal Welfare, Neubiberg/Munich, Germany; Transinsight GmbH, Dresden, Germany; and German Federal Institute for Risk Assessment (BfR), Center for Alternative Methods to Animal Experiments - ZEBET, Berlin, Germany. Go3R is on-line since April 2008 under www.Go3R.org.

The general problem of identifying ontology terms in text will be discussed emphasising specifically anatomical and developmental terminology as well as concepts relevant to the search for animal testing alternatives. Therefore this chapter mentions a number of semantic search applications using ontologies, whose creation and maintenance is supported by the ontology learning methods developed in this thesis. It will be described how the technology developed for GoPubMed (Doms and Schroeder, 2005; Doms, 2009; Dietze et al., 2008a,b) has been successfully transferred to other knowledge domains, in particular the search for literature on animal testing alternatives as mentioned above. In this regard, three additional projects will be presented:

GoPubMed *is an ontology-based search engine using background knowledge to help answering biomedical questions. GoPubMed retrieves PubMed abstracts for your search query, detects Gene Ontology and MeSH terms in the abstracts and allows the user to browse the search results by exploring the ontologies and displaying only papers mentioning selected terms, their synonyms, or descendant in the ontology. GoPubMed is online since 2005 under www.gopubmed.org.*

MousePubMed *is a research project to accommodate the specifics of anatomy, which works with genes, tissues, and developmental stages as used in the Edinburgh Mouse Atlas (Baldock et al., 2003). MousePubMed's automated annotation of PubMed abstracts was evaluated against the handcurated annotations of the Edinburgh Mouse Atlas.*

LMOPubMed *is an ontology-based search engine which has been developed in collaboration with Unilever Research (Colworth, UK) and Transinsight GmbH. LMOPubMed indexes documents using terms from the newly created Lipoprotein Metabolism Ontology (LMO). The goal of LMOPubMed is to categorise lipids with respect to risk factors for diseases and ethnic specifics.*

8.1 Improving literature searches with semantic search technologies

The scientific and technical basis of a semantic search engine is its underlying specific expert knowledge. In semantic search engines as described here, this specific expert knowledge is captured within an ontology. An ontology is an extensive and detailed network of — also hierarchically — grouped concepts. If available, concepts also contain information about synonymous terminology. The ontology specifies the unambiguous meaning of relevant terms and depicts the complex relationships existing between them. With the help of such an ontology, the topic and content of any document can be semantically determined by the mapping of the unique pattern of concepts and terms utilised in it.

Our, semantic search engines present search results in a structured form, accompanied by an “intelligent table of contents” provided by the taxonomic backbone of the ontology. The searcher can use this table of contents to navigate through the search result and quickly extract those pieces of information that are relevant to him.

Furthermore, semantic search engines understand what the user is searching for and even retrieve information that has not explicitly been sought for. The following example of the question “which enzyme inhibits aspirin?” illustrates the difference between a conventional search in PubMed of the US National Library of Medicine and the National Libraries of Health and a knowledge-based search. In this example, the knowledge-based search was performed with GoPubMed (Doms and Schroeder, 2005; Doms, 2009), a general biomedical search engine, which already uses knowledge-based semantic technologies. In PubMed, the search query “aspirin” results in a retrieval of 41,257 documents¹, which are presented in the form of a long list. Whereas GoPubMed delivers the same large amount of retrievals, they are accompanied by an “intelligent table of contents” (Figure 8.1). By scrolling through this table of contents, the vast pool of documents can already be reduced to 30 documents with two clicks by filtering the search result with the category “Chemicals and Drugs” (#1 in Figure 8.1) and the term “Cyclooxygenase 2” (#2). The first relevant document of this sub-list of 30 of the total of 41,257 documents would have been listed on position No. 410 of the PubMed list (#3). The third document (#4) of the mentioned sub-list then provides the answer that PGHS-2, a synonym to Cyclooxygenase 2, is inhibited by aspirin. In conclusion, the pre-sorting of documents provided by the knowledge-based search engine enables a speedy, precise and goal-oriented answering of the respective scientific question sought for.

Related work on semantic search in biology and medicine The enterprise Vivisimo (Taylor, 2007) takes the same approach as GoPubMed by displaying a categorisation of search results in a tree structure. With the product “Clusty”, Vivisimo offers a meta search engine which groups similar results together into clusters. Still, the approach differs from GoPubMed as Clusty does not use a controlled vocabulary to index documents. Representative labels are identified from each cluster. Another example of an ontology-based search engine is Textpresso (Müller et al., 2004) which was built for scientific literature on *C. elegans*, a well-known model organism in biology, and selected others domains. The documents are indexed with biolog-

¹ result as of January 2009

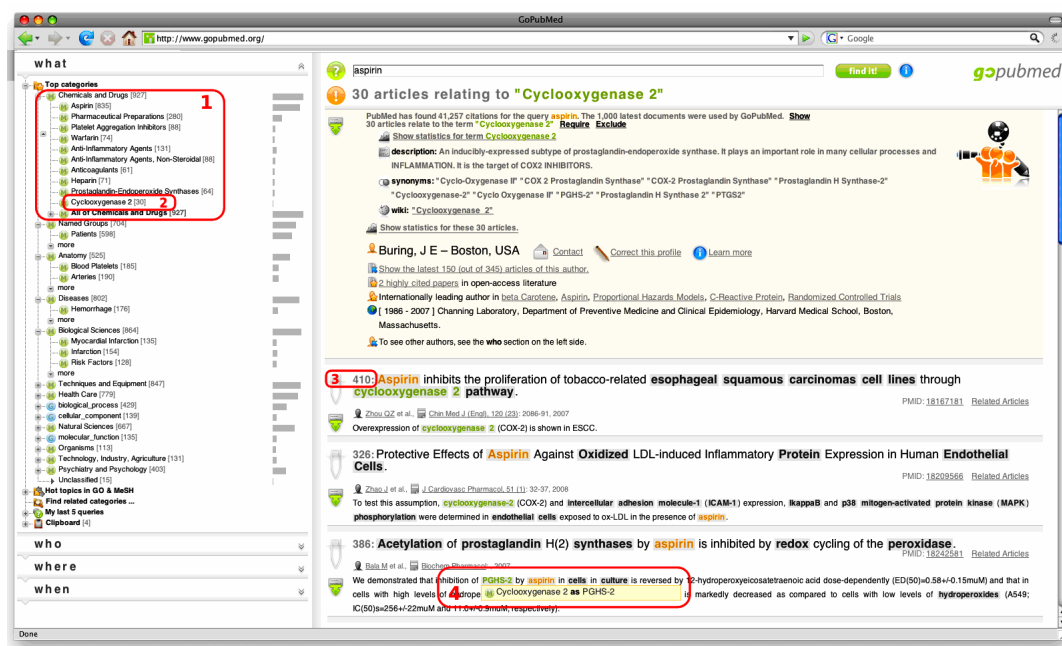


Fig. 8.1. Semantic search in GoPubMed to answer the question which enzyme inhibits aspirin?

ical concepts and relations based on a flat list of 101 concepts. EBIMed (Rebholz-Schuhmann et al., 2007), a tool developed by the European Bioinformatics Institute, identifies associations between UniProt protein/gene names, Gene Ontology annotations as well as Drugs and Species. Here, results are presented in chart form; and sentences supporting the associations are cited. The knowledge retrieval tool iHOP (Hoffmann and Valencia, 2004) hyperlinks Medline/PubMed articles via protein or gene names. The user can query for a gene or protein and receives answers in form of sentences suggesting interactions with another gene/protein. The sentences are marked up with MeSH terms as well. In Doms et al. (2006) we described this use of ontologies and text mining for semantic web applications.

Importance of literature search in biology and medicine A primary source of data is stored in special purpose databases. Queries across disparate databases are required to exploit available data. However, a lot of data is not yet stored in such a structured form. This is due to two main reasons. For one, there is no immediate interest for researchers to submit their findings to (one or more) relevant databases, as scientific publications as the main instrument for making information accessible and gaining reputation. The second reason comes with the necessary process of manual curation of database entries and annotation to maintain a certain quality standard. Another resource of data are aforementioned scientific publications themselves. Fairly often, these provide insight into more recent findings than databases. In addition, more information can be found in texts, such as, background knowledge, descriptions of experimental settings, etc., showing broader context as well as in-depth details. Natural language often is more suitable to express facts than the structured form of any database. Moreover, many annotations in databases come in the form of free text, for instance functions and diseases in UniProt, or

phenotypes in the Mouse Genome Informatics database (MGI). This shows that scientific publications and other textual descriptions present important resources to be considered when searching for complete information. In the following sections it will be described, how ontological terms can be found in text.

Text mining

In biomedical text mining, researchers use techniques from natural language processing, information retrieval, and machine learning to extract desired information from text (Jensen et al., 2006). Even when the concepts to extract are available in a structured form, such as a controlled vocabulary or ontology, finding them in free text is not always an easy task. For instance, recent assessment for extracting Gene Ontology terms revealed precision and recall above 70% (Table 8.1). The difficulty of automating manual annotation is evident from the fact that only as few as 15% of manually annotated terms appear literally in the associated abstracts.

System	Precision	Recall	Benchmark Size
FiGO (Couto et al., 2005)	28.7%	N/A	301
MeKE (Chiang and Yu, 2004)	49.8%	N/A	125
LTA (Doms, 2004)	35.8%	85.7%	18356
ConceptRecognition (Doms, 2009)	79.9%	72.7%	18356

Table 8.1. Overview over Gene Ontology term recognition algorithms. The algorithm Local Term Alignment (LTA) and ConceptRecognition which performs best are used in Go3R.

Ad-hoc variations of names To begin with, terms in vocabularies and labels of concepts in ontologies appear in many, slight or severe, variations in natural language texts.

- orthographic: *IFN gamma*, *Ifn- γ*
- morphological: *Fas ligand*, *Fas ligands*
- lexical: *hepatitic leukaemia*, *liver leukemia*
- structural: *cancer in humans*, *human cancers*
- acronyms/abbreviations: *MS*, *Nf2*
- synonyms: *neoplasm*, *tumor*, *cancer*, *carcinoma*
- paragrammatical phenomena/typographical errors: *cerevisae*, *nucleotid*

Some of the terms encountered in texts are rather ad-hoc creations, which cannot be found in any term lists (compare also Section 2.3.1).

Synonymy of ontological terms As mentioned before, terms in a vocabulary or ontology might not appear literally in a text, but authors rather use synonyms for the same concept. First of all, this complicates proper searches: When searching for “digestive vacuole”, results should also contain texts that mention “phagolysosome”; mentions of “ligand” refer to the concept “binding”; an “entry into host” might occur as an “invasion of host”. In the Plant ontology for example, many synonyms exist for the same structure in different species. “Inflorescence” is referred to as “panicle” in rice, and as “cob” in sorghum, and “spike” in wheat. It has to be

noted that there is also intra-ontology synonymy. “eye” in AnoBase can refer to the “eye spot” or the “adult compound eye”. In a similar manner, the Edinburgh Mouse Atlas contains unspecific mentions such as “cavity” or “body” for the mouse.

Ambiguity of ontological terms Terms can have a very specific meaning in biomedical research, but mean other things in other domains. Examples are “development”, “envelope”, “spindle”, “transport”, and “host”. Protein names such as “dreadlocks”, “multiple sclerosis” or “the” that resemble common names, diseases, or common English words are especially hard to disambiguate. The same problems arise from drug names like “Trial” or “Act”. Table 8.2 lists some anatomical terms that have other meanings in different domains. Especially where cross-ontology or cross-database queries are needed, one has to consider ambiguity, for instance when applied to different organisms: “gametogenesis” (sexual reproduction) in plants is different from “gametogenesis” in metazoans.

Stemming and missing words Some aspects for finding terms in text refer to the actual processing of natural language and appear rather technical. Very often, words will appear in different forms, such as “binding” and “binds”. These refer to the same concept, which can be solved by resolving words to their stem (“bind”). However, the analogous reduction of “dimerisation” to “dimer” is more questionable. The former talks about the process, the latter about the result. A similar example is “organisation”, where a transformation into “organ” is invalid.

Texts contain additional words that are missing in the ontological term. This happens, for instance, when a text contains further explanations that describe findings in more detail. An example is “tyrosine phosphorylation of a recently identified STAT family member” that should match the ontology term “tyrosine phosphorylation of STAT protein”. In general, matching is allowed to ignore words such as “of”, “a”, “that”, “activity”, but obviously not “STAT”. Additional background information on term variations is needed to know that a “family member” can refer to a protein.

Formatting of terms represents another source for potential matching errors. Terms in an ontology contain commas, dashes, brackets, etc., which require special treatment. For “thioredoxin-disulfide” the dash can be dropped, for “hydrolase activity, acting on ester bonds” the clause after the comma is important, but unlikely to appear as such in text. Terms containing additions such as “(sensu Insecta)” contain important contextual information, but are also less likely to appear in text.

Ontologies and text mining

Three main key dimensions of ontologies have been defined by Uschold: formality, purpose, and subject matter (Uschold and Grüninger, 1996). The degree of formality by which a vocabulary is created and meaning is specified varies among different ontologies. The purpose refers to the intended use of an ontology. Domain ontologies (such as medical or anatomical), problem solving ontologies, and representation ontologies comprise examples for different subject matters an ontology is characterising.

In contrast to ontologies designed primarily for annotating biological objects, there is a clear distinction to ontologies designed for text mining. This distinction and its impact on text mining strategies as well as on the redesign of dedicated

ontologies will be described. In the case of a text mining ontology, compromises must be made on the relationships and on the labels used. Labels need to be descriptive and they or associated synonyms must be used in text. The ontology does not need to be very formal in terms of containing many different relationships between terms (such as “derives from”, “causes”, “part of”, etc.) or of distinguishing between “classes” and “instances”. It should be constructed in a way, that it is possible to obtain a structured vocabulary with only one type of directed relationship defining a hierarchy, i.e. “is_a” relationships or simply parent child relationships. In general, there has to be a compromise to obtain a correct ontology with valid relations and still get the best possible results from text mining. The most prominent topics considering ontology design for text mining are the following.

Term overlaps Some concepts can overlap in their labels or synonyms: in many cases there is a difference between what authors write and what they actually mean to express. Unfortunately, researchers do not have strict and formal ontologies or nomenclatures in their minds when composing a scientific article; in most of the cases they might use parent terms to refer to a child term, or vice-versa. For example, many people are treating the MeSH terms “cardiovascular disease” and “coronary artery disease (CHD, CAD)” the same, although the latter is a child of the first.

Descriptive labels In most of the cases, the labels in an annotation ontology cannot be used for text mining, usually due to their explanatory nature. For example, it is unlikely that the Gene Ontology term “cell wall (sensu Gram-negative bacteria)” will appear as such in text. Terms like “positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism” and “dosage compensation, by inactivation of X chromosome” are almost complete sentences and are also unlikely to be found as such in text.

Ambiguity results either from identical abbreviations for different terms, or tokens that can refer to terms that might be or may not be of our interest. An example of an ambiguous abbreviation is “CAM” that can stand for “constitutively active mutants”, “cell adhesion molecule”, or “complementary alternative medicine”. The second category of ambiguities — and the most difficult to handle — is that of terms that (in the context of anatomy) can refer to different species. An example of such ambiguities is “embryo”, which can be a chicken, mouse, human, or even zebrafish embryo. Therefore, if one is interested in the different developmental stages of the mouse embryo nervous system, it is required to retrieve articles focusing on studies on mouse embryos only. If the term “embryo” is inserted in the Mouse Anatomy ontology as such, then the search engine will return articles on all kinds of embryos. If the term “mouse embryo” is inserted in the ontology, the number of articles retrieved will not be the real number of articles mentioning the term “mouse embryo”, since not all of them will mention the term as such. A similar example is that of organs/tissues common to different species, such as “eye” or “lens”. This kind of ambiguity is especially difficult as the species is often not explicitly mentioned.

Generic and specific labels When using the ontology for text mining in a specific biomedical sub-domain (anatomy, disease, glucose metabolism, etc.), the ontological

concepts must be specific for that domain. The articles retrieved must be anatomy-specific, disease-specific, or glucose-metabolism-specific. Therefore, a vocabulary is needed, which is specific enough to distinguish between relevant and irrelevant articles, but general enough to not exclude potentially relevant articles. If the concepts are too generic, they could be referring to many other domains. For example, during the design of a glucose-metabolism ontology, information on kinetics might need to be included. “Kinetics” as such is too generic to be used as a term, as it can refer to different kinds of kinetics (kinetics of phase transition, hydrolysis kinetics, kinetics of equilibrium reactions). On the other hand, the term “glucose kinetics” might be too specific, as it might seldom appear as such in a text. The decision on which terms should be used in the ontology ideally should only be made after exhaustive searches with different variations of terms.

Rules for text mining ontologies Some simple rules can be derived from all these observations, which can be used for the (re-)design of ontologies when they should serve as resources for text mining applications.

- Avoid descriptive labels and synonyms: labels should be likely to appear in texts as such – avoid “and”, “of” and the like;
- Avoid improper spelling variations: capitalisation, noun plural forms, verb flexions;
- Use common names as labels or include them as synonyms;
- Add structural and lexical variations wherever possible;
- Keep the nomenclature consistent, precede terms with superstructure name;
- Use different representations of a concept in the ontology.

For a proper extraction of terms and subsequent term disambiguation in case of homonyms, the occurrence of parents helps to decide on the exact term. Especially in anatomical ontologies, terms can have multiple representations, such multiple hierarchies should also be reflected by the ontology. Examples are spatial and systemic representations of a tissue — “lung” is a “body part”, and also a specific “organ system”. Depending on the context in which “brain” is found, parent terms below “head” might not be found in the text at all, but rather terms related to “organ system.” An ontology should therefore cover at least the most likely paths to subsume a tissue.

All of the above problems mean that extracting terms from literature will not be error-free. However, despite all of these problems, ontology-based literature with text-mining can answer questions as posed in the introduction. Next, the three search engines, GoPubMed, MousePubMed, LMOPubMed, and finally Go3R will be introduced and it will be illustrated how they have been developed and how they help to answer questions.

8.2 GoPubMed

GoPubMed (Doms and Schroeder, 2005) and MousePubMed (Wächter et al., 2007), which is discussed in the next section, index articles provided by PubMed with ontology terms from GO, MeSH, and Mouse anatomy/development, respectively. As an example consider Figure 8.2 and Figure 8.3 which show screenshots of GoPubMed when queried for Pax6. The key difference to a classical search is that all the documents are annotated with terms from the domain specific ontology. Therefore, the user interface shows ontological information on the left and the documents on the right side. Beside the complete hierarchy of relevant terms found in documents mentioning the given keywords, a list of frequently occurring terms is placed above. Clicking on any of these terms reduces the result set and allows users to quickly filter large result sets to the necessary documents needed to answer their question.

Answering questions The ontological background knowledge can serve to answer questions with such tools. Consider for example a researcher interested in the Pax6 gene. He/she might have the following questions:

- *Which processes is Pax6 involved in?*
- *Which diseases is Pax6 involved in?*
- *At which developmental stages is Pax6 active in mice?*

Literature holds answers to these questions, but a classical literature search cannot answer the questions directly, as articles will not mention gene, disease or process, but rather specific instances such as Pax6, Aniridia, or eye development. Since ontologies contain knowledge that Pax6 is a gene, Aniridia is a disease, and eye development is a process, they can help to answer questions as the following:

- *Which processes is Pax6 involved in?* A query in GoPubMed for Pax6 shows that the most frequent processes mentioned are “gene expression”, “regulation of gene expression”, and “eye development” (Figure 8.2). Opening the development branch reveals the processes of brain and eye development as well as pancreas development. Indeed the corresponding articles support this essential role of Pax6 as transcription factor and master control gene in development of eye, brain and pancreas (Hsieh and Yang, 2009).
- *Which diseases is Pax6 involved in?* A query in GoPubMed for Pax6 shows that the most frequent disease mentioned is aniridia. Hovering the mouse over the term gives an explanation that it is “a congenital abnormality in which there is only a rudimentary iris”. This is due to the failure of the optic cup to grow. Aniridia also occurs in a hereditary form, usually autosomal dominant.” A click on aniridia shows articles mentioning both the disease and the gene such as for example Castori et al. (2009), which confirm the answer as “Aniridia is a developmental disorder of the eye due to heterozygous mutations in PAX6.” (Figure 8.3).

Indeed, Pax6 is the most researched gene of the family of Pax genes and appears throughout the literature as a “master control” gene for the development of eyes and is of medical importance because heterozygous mutants produce a wide spectrum of ocular defects such as aniridia in humans. The question “*At which developmental*

my search

Home Go3R GoGene GoWeb Help

Pax6 eye development[go] **find it**

264 of 1,473 documents analyzed

top author

statistics

documents

Dynamic Pax6 expression during the neurogenic cell cycle influences proliferation and cell fate choices of retinal progenitors.
 Hsieh, Y.W. et al.
 Journal: *Neural Dev.* Vol. 4 (1): 32, 2009
 Here, we examine the dynamic changes of Pax6 expression among chicken retinal progenitors as they progress through the neurogenic cell cycle, and determine the effects of altered Pax6 levels on retinogenesis.
 PMID: 19686589 Related Articles

Targeted deletion of Dicer disrupts lens morphogenesis, corneal epithelium stratification, and whole eye development.
 Li, Y. et al.
 Journal: *Dev Dyn.* Vol. 238 (9): 2388-400, 2009
 We studied the roles of Dicer and miRNAs in eye development by conditionally deleting the Dicer gene in the mouse lens and corneal epithelium.
 Affiliation: Laboratory of Molecular and Developmental Biology, National Eye Institute, National Institutes of Health, Bethesda, Maryland.
 PMID: 19681134 Related Articles

Stage-dependent modes of Pax6-Sox2 epistasis regulate lens development and eye morphogenesis.
 Smith, A.N. et al.
 Journal: *Development.* Vol. 136 (17): 2977-85, 2009
 The cooperative activity of Sox2 and Pax6 is illustrated by the dramatic failure of lens and eye development in presumptive lens conditional, compound Sox2, Pax6 heterozygotes.
 Affiliation: Division of Pediatric Ophthalmology, Cincinnati Children's Hospital Research Foundation, Cincinnati, OH 45229, USA.
 PMID: 19668824 Related Articles

Fig. 8.2. GoPubMed query for “Pax6”. On the left, the knowledge base (intelligent table of contents) with frequent terms by category and all relevant terms is shown. The third most frequently mentioned biological process is eye development. Clicking the term and retrieving the articles mentioning eye development. The top most result mentions that “the effects of altered Pax6 levels on retinogenesis”. Here “retinogenesis” is a synonym for “retina morphogenesis in camera-type eye” which again is a descendant of “eye development” in the Gene Ontology.

my search

Home Go3R GoGene GoWeb Help

Pax6 Aniridia[mesh] **find it**

196 of 1,473 documents analyzed

top author

statistics

documents

Darier disease, multiple bone cysts, and aniridia due to double de novo heterozygous mutations in ATP2A2 and PAX6.
 Castori, M. et al.
 Journal: *Am J Med Genet A.* Vol. 149A (8): 1768-72, 2009
 Aniridia is a developmental disorder of the eye due to heterozygous mutations in PAX6.
 Affiliation: Experimental Medicine Department, 'Sapienza'-University of Rome, San Camillo-Forlanini Hospital, Italy.
 mcastori@scamilioforlanini.mi.it
 PMID: 19610080 Related Articles

A novel missense mutation (Leu46Val) of PAX6 found in an autistic patient.
 Maeakawa, M. et al.
 Journal: *Neurosci Lett.* Vol. 462 (3): 267-71, 2009
 Mutations in PAX6 are responsible for eye abnormalities including aniridia, and it is also known that some PAX6 mutations result in autism with incomplete penetrance.
 Affiliation: Laboratory for Molecular Psychiatry, RIKEN Brain Science Institute, Wako, Saitama, Japan. mmaekawa@brain.riken.jp
 PMID: 19607881 Related Articles

Selective cortical layering abnormalities and behavioral deficits in cortex-specific Pax6 knock-out mice.
 Tuoc, T.C. et al.
 Journal: *J Neurosci.* Vol. 29 (26): 8335-49, 2009
 Because a majority of the morphological and behavior disabilities of the Pax6 mutant mice parallel abnormalities reported for aniridia patients, a condition caused by PAX6 haploinsufficiency, the Pax6 conditional mutant mice generated here represent a valuable genetic tool to understand how the developmental cortical disruption can lead to a human behavior abnormality.
 Affiliation: Department of Molecular Cell Biology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany.
 PMID: 19571125 Related Articles

Fig. 8.3. GoPubMed query for “Pax6”. On the left, the knowledge base (intelligent table of contents) with frequent terms by category and all relevant terms is shown. The most frequently mentioned disease is aniridia. Clicking the term and retrieving the articles mentioning aniridia confirms that Pax6 is involved in aniridia. The top most result mentions that “Aniridia is a developmental disorder of the eye due to heterozygous mutations in PAX6.”

stages is *Pax6* active in mice?” can be answered with MousePubMed introduced in the following section.

We can now further check in GoPubMed whether aniridia is a “hot topic” (Figure 8.4) and who the most active authors publishing on aniridia are (Figure 8.5). It turns out that V. van Heyningen is the number one publishing author also having the most collaborations, as shown on the co-authorship network.

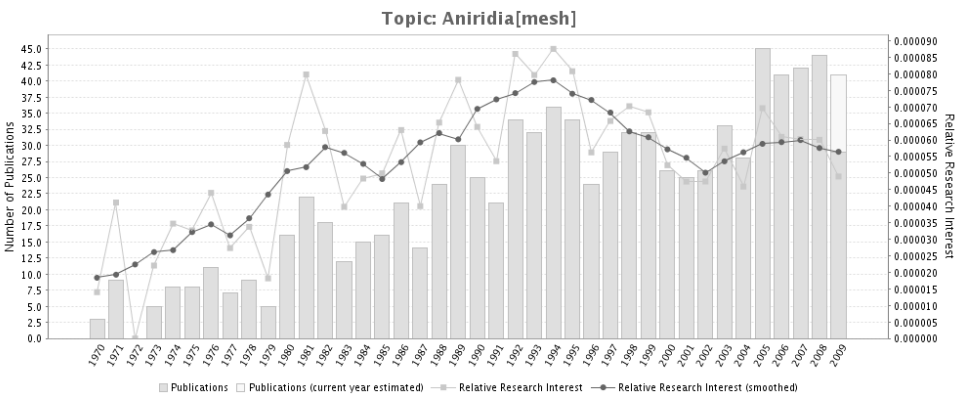


Fig. 8.4. GoPubMed. Chart showing the absolute and relative number of publications about the disease “Aniridia” over time.

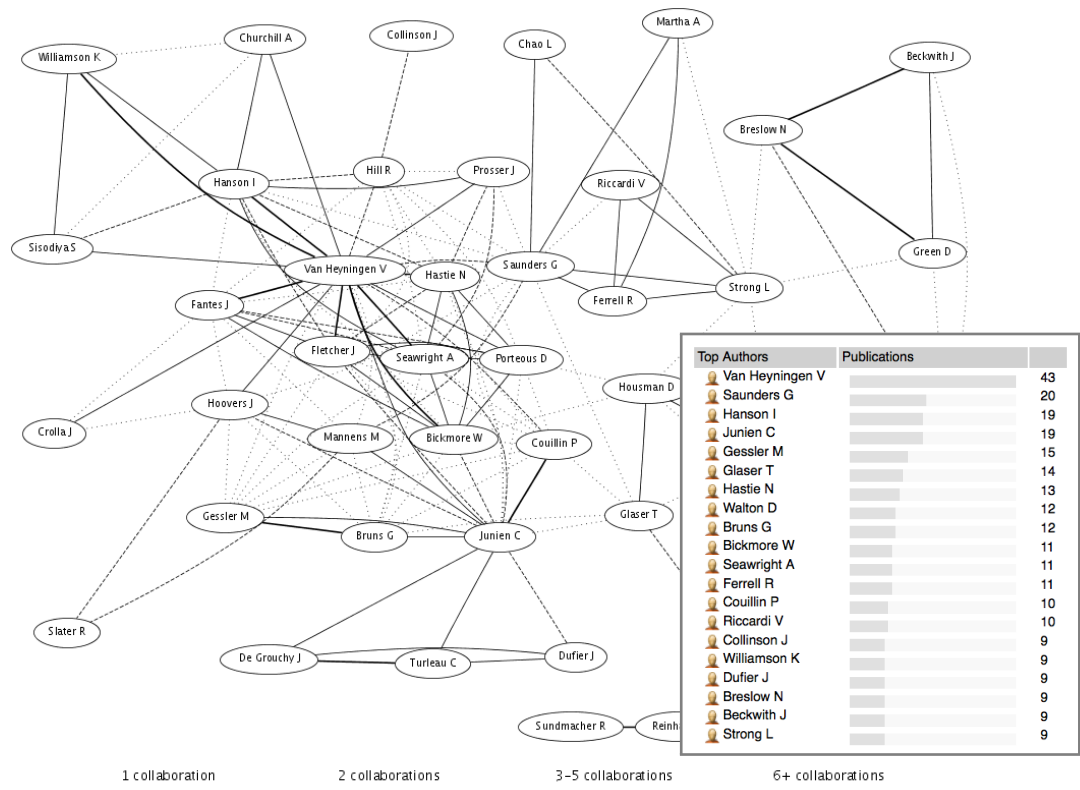


Fig. 8.5. GoPubMed. Part of the co-authorship network for “Aniridia” in GoPubMed and author statistics showing V. van Heyningen as the author most active in this area.

8.3 MousePubMed

MousePubMed is a prototype for an ontology-based search engine which was created to interrelate genes and tissues with specific developmental stages in mouse development.

Reference:

Wächter, T., Alexopoulou, D., Dietze, H., Hakenberg, J., and Schroeder, M. (2007). *Anatomy Ontologies for Bioinformatics, Principles and Practice*, volume 6, chapter Searching Biomedical Literature with Anatomy Ontologies. Springer Computational Biology.

8.3.1 Introduction

Ontology-based literature search evaluated against the Edinburgh Mouse Atlas

Many ontologies and vocabularies have been designed to annotate genes and gene products based on evidence from literature. They are also useful to search literature systematically. GoPubMed presented before is such an ontology-based literature search engine. It allows users to explore PubMed search results with hierarchical vocabularies Gene Ontology (GO) and MeSH. MousePubMed, an adaption of GoPubMed, is a prototype of a web-based search engine that is able to establish a link between anatomical concepts, stages in mouse development, and known genes with impact on mouse development. It provides the search result automatically linked to an “intelligent table of contents” and disambiguates anatomical concepts according to their phase in development. The Edinburgh Mouse Atlas with genes, tissues, and developmental stages was integrated in the search engine. More specifically MousePubMed uses vocabularies for mouse anatomy (EMAP), human anatomy (EHDA), mouse genes (from EMAGE), and mouse developmental stages (Theiler) as resources. Specialised text mining algorithms have been developed to match highly ambiguous anatomy terms and Theiler stages. Coming back to the third question on Pax6 in the previous section, which can be answered with MousePubMed:

- *At which developmental stages is Pax6 active in mice?* A query in MousePubMed for Pax6 shows that Theiler stages up to 14 (9 dpc, days post conception) are frequently mentioned supporting Pax6’s role in early development. Clicking on a stage reveals e.g. the statement “In the early development of the vertebrate eye, Pax6 is required for...” in Azuma et al. (2005)

To demonstrate its usefulness, GoPubMed has been evaluated against the Mouse Atlas. For nearly 1500 genes and over 10.000 triples of gene, tissue and stage, it was possible to reconstruct with MousePubMed 37% of genes, 31% of gene-tissue associations, and 13% of gene-tissue-stage associations on the basis of PubMed abstracts. These figures are encouraging as only abstracts are used. Before discussing this evaluation the newly developed matching algorithm will be introduced.

Extracting gene names, anatomy terms, and developmental stages

Ontology based text mining is not restricted to finding words or word groups in texts. The structure of the ontology can be used to state the relation between a term

The screenshot displays the MousePubMed interface. On the left, a 'Hierarchy of relevant terms' sidebar shows a tree structure starting with 'mouse AND Pax6' and branching into various biological categories like 'Embryo', 'Organ system', and 'Tail'. The main content area on the right shows search results for the query 'mouse AND Pax6' and term 'TS_14'. It includes a 'GO terms' section with highlighted keywords 'mice, mouse, Pax6', a 'Textmined' section with 31 terms, and a list of articles. The first article is highlighted: 'COUP-TFs regulate eye development by controlling factors essential for optic vesicle morphogenesis.' by Tang K, Xie X, Park JI, Jamrich M, Tsai S, Tsai MJ.

Fig. 8.6. Screenshot of MousePubMed. The left side shows the ontology which comprises tissues, genes, and developmental stages (Theiler stages). Documents are assigned to the nodes of the ontology using different text-mining algorithms. Tissues are matched and disambiguated, and developmental stages are found according to manual created patterns. The right side shows the documents itself with the terms highlighted.

and a document by finding the children of the term. This task is reasonably well solvable for the Gene Ontology where its term labels are self-descriptive. Many terms in GO are contained in their child terms (Ogren et al., 2004). As an example, the term “envelope” is refined into “organelle envelope” and further to “organelle envelope lumen”. The ontology for the Abstract Mouse contains anatomical concepts in the mouse embryo at different embryonic developmental stages. The vocabulary is used to annotate images of mouse embryos. It unifies the vocabulary needed to describe the different parts throughout 26 Theiler stages. Concepts like organs or body parts are further refined into tissue types, unspecific loci such as “cavities”, “left”, “upper”, as well as general terms such as “node” or “skin”. Considering only the textual labels, one cannot distinguish between the different ontological concepts. For example, “chorion” has the children “mesoderm”, “ectoderm” and “mesenchyme”. “Amnion” and “yolk sac” have children sharing the same labels. Searching for documents related to “chorion” will retrieve very similar document sets to searching for “amnion”, only because the documents mention “mesoderm”, in this case with meaning “mesoderm specific to amnion”. Different anatomical concepts share the same term label. For instance, there exist 171 individuals with label “epithelium”. These all refer to different body parts at a specific stage in development.

Ontology-based text mining relies on the assumption that unique or similar types of directed non-cyclic relationships exist, which can be unified in the hierarchical

relationships creating a taxonomy. This assumption does not hold for the Abstract Mouse ontology. There does not always exist a path to the common root supported by only one type of hierarchical relationships. Therefore, in our analysis, a document is annotated with a term from the Abstract Mouse ontology taking the term label and its synonymous labels into account. In the Abstract Mouse Ontology the term labels follow various creation patterns. Sometimes a child term contains information of the parent term (for example, “cavities” has the child “amniotic cavity”). In other cases a term like “umbilical vein” has the children “left” and “right”, rather than “left umbilical vein” and “right umbilical vein”, respectively. These short and common sense labels make the text annotations arbitrary.

For our experiments we slightly adapted the ontology. For the terms “left”, “right”, “upper”, “lower”, “common”, “anterior” and “posterior” we expanded the term labels with its parents labels. “Eyelids” thus became “upper eyelids” and “lower eyelids”, for instance, and we removed the children terms “upper” and “lower” accordingly. To distinguish between common terms such as “skin” occurring — for instance, for different organs — the matching algorithm took text annotations for ancestor terms into account. Terms with the same label were grouped according to the number of text annotations for their ancestors in the same document. Only annotations of the top ranked group were confirmed. Figure 8.7 shows an example for the term “skin”. There were multiple possibilities to resolve this term to a specific tissue. Only when a parental term (shoulder, upper arm, etc.) was found, the mention was annotated with the specific skin.

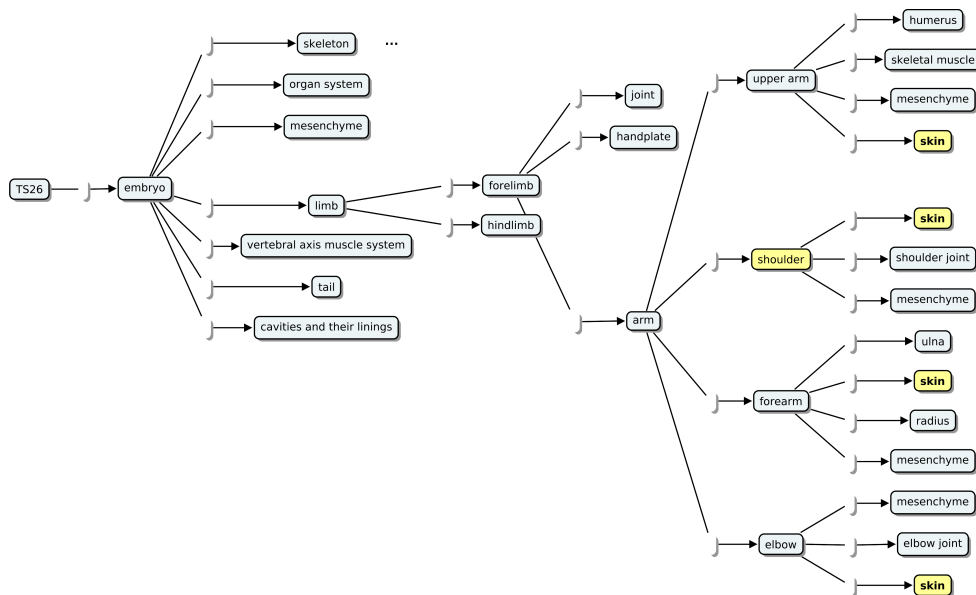


Fig. 8.7. Excerpt of the anatomy ontology used in MousePubMed showing the different types of **skin**. Occurrences of the term “skin” (yellow concept nodes) in a text were resolved using the hierarchical dependencies. Only when a parental node was also found, for instance, “shoulder”, we annotated the text with “skin.”

As Table 8.3 shows, nine PubMed abstracts contained the full information as stated by Thut et al., mentioning gene, tissue, and specific stages (days). For most

Term	Alternative meaning
rod	common English
iris	species: plant; common English
axis	species: deer; common English
chin	common English
beak	common English
pons	protein: Serum paraoxonase/arylesterase 1 (PON)
penis	protein: Penicillinase repressor (penI)
sigma	common English/Greek
patella	species: limpet
cicatrix	disease: scar
nephrons	drug: bronchodilator (Nephron)
hemocytes	drug: iron supplement (Hemocyte)
chondrocytes	drug: cartilage cells for implantation
hippocampus	species: seahorse

Table 8.2. Anatomical terms with different meanings in other knowledge domains. Some misinterpretations occur only when certain spelling variations are allowed, for instance, ignored capitalisation or plural forms.

Gene	Tissue	Stage	PubMedID
Sparc	retina, RPE, eye	E4.5, E5, E10, E14, E17	9367648
Sparc	lens	embryonic day (E)14	16303962
Stat3	retina, RPE, eye	-no specific stage-	12634107
Stat3	lens	E10.5	14978477
Pedf	RPE	-no specific stage-	7623128
Pedf	retina	E14.5, 18.5	12447163
Runx1	inner retina	embryonic day 13.5	16026391
Col15a1	conjunctiva, cornea	E10.5-18.5	14752666
Otx2	outer retina	-no specific stage-	15978261
Edn1	retina	-no stage-	11413193
IGF-II	eye, cornea, retina, scleral cells	E14	2560708
Wnt7b	anterior eye, cornea, optic cup, iris	-no specific stage-	16258938
CDH2	—	-no stage-	9210582
—	lens	-no stage-	9211469
Col9a1	eye, lens vesicle, neural retina, ciliary epithelial cells, cornea	13.5, 16.5-18.5 d.p.c.	8305707
Tgfb2	cornea, lens, stroma	-no specific stage-	11784073
Thra	retina	-no specific stage-	9412494
BMP4	retina	E5	17050724
Bmp4	optic vesicle, lens	-no specific stage-	15558471
BMP4	lens, optic vesicle	-no specific stage-	9851982
—	eyes	N/A	15902435
Sox1/2	lens	-no stage-	15902435
—	retina, eye axis	E2, E3, E5	15113840
Notch1	eye	-no specific stage-	11731257
Notch2	eye	-no specific stage-	11171333

Table 8.3. Expression patterns identified by MousePubMed in articles derived from Thut et al. (2001). Often, an abstract does not mention a (specific) developmental stage; —: MousePubMed did not find this particular fact; otherwise: facts as identified by MousePubMed. Given are only tissues related to the murine eye.

cases, however, not all data were contained in one single abstract. In three cases, we were not able to automatically spot the gene name (left column), in all cases this was due to synonyms lacking in EMAP and MGI. Note that the assessment of recognising genes was based only on genes mentioned in EMAGE. The tissue could be found in almost all of the cases; from most abstracts, even the specific part of the eye could be extracted.

Finding gene names in documents is done using exact matching against gene names contained in EMAGE. We enriched this set using additional names and synonyms for each gene taken from the MGI database². We tested all 1437 genes mentioned in EMAGE for their annotations with tissues and Theiler stages in PubMed.

We analysed 123,074 abstracts retrieved from PubMed with the query mouse AND development. This amounted to approximately 0.7% of all documents listed in PubMed. Based on the document annotations with ontology terms, we issued in total 36,358 statements on relations between genes, tissue and developmental stages, which we extracted from EMAP/EMAGE. Cases with multiple Theiler stages from EMAP were split into separate statements. We evaluated the tissues mentioned using EMAP's Abstract Mouse ontology and the anatomy part of MeSH. For path descriptions like "embryo.ectoderm" in EMAP we required the matching document to be annotated with the terms "embryo" and "ectoderm". For MeSH, as in GoPubMed, we also included descending terms. A document was annotated with the term "embryo" if annotations for its descendants, for example, "germ layers" or its children "ectoderm", "endoderm" or "mesoderm", were found.

To find mentions of Theiler stages in texts, it was not enough to search for them directly, as they seldom occur as such in abstracts ("Theiler stage 12", "TS12", etc.). We therefore compiled a set of regular expressions based on two main notions, the mentioning of embryonic days (E) and of days post coitum (dpc). These expression had to capture occurrences like

- "embryonic day 10.5",
- "day 9 mouse embryos",
- "between E3.5 (E = embryonic day) and E8.5",
- "12.5 days post coitum", and also
- "7.5-13.5 days post-conception."

As mentions of Theiler stages do not often occur, but rather general time spans are given ("early embryonic development"), we decided to assign Theiler stages one to 14 to "early development", and stages 20 to 27 to "late development," respectively. Every mention of an "early developmental stage" thus was treated as a match for stages one through 14. Both assignment were based on statements found in PubMed relating days to general time spans.

8.3.2 Experiment designs

To assess the potential of ontology-based literature searches, we designed two experimental scenarios. For the first, we manually collected two sets of queries and detailed answers. For the second scenario, we evaluated the complete EMAP/EMAGE

² See <http://www.informatics.jax.org>.

data. Using the methodology described in the previous section, we tried to find textual evidences for all sets in PubMed. This means that we searched PubMed for abstracts that shared annotations for each collected triple consisting of a gene, tissue, and Theiler stage.

Manually curated test set

We first selected set of questions manually to study results in detail. The idea was to send simple keyword queries to MousePubMed, asking for mouse abstracts that discuss a certain tissue and embryonic day. MousePubMed should then identify all genes mentioned in the top-ranking abstracts. Questions and retrieved answers were as follows.

- *Which genes play a role in the development of the nervous system in Theiler stage 14?* A query for “mouse development nervous system 9 dpc” finds the genes Adamts9, Hoxb4, Otx3, and EphA4 within the first eight abstracts³. In addition, the genes EphA2, A3, A7, B1, B2, and B4 are found, which are not yet annotated in the EMAGE database.
- *Which genes play a role in sex differentiation during murine embryo development?* A corresponding query for “mouse sex 10 dpc” results in a set of eight genes within the first fifteen abstracts: Fgf9, Asx11, Sry, Sox9, Usp9x, Maestro/Mro, Wt1, Amh1 and Fra1⁴. Only half of the genes can be found in EMAGE so far.
- *Which genes play a role in the development of the murine embryonic liver?* A query for “mouse liver development” results in a set of several genes, most of which can be found in EMAGE as well: Shc, Pxn, Grb2, PEST/Pcnp, GATA6, HNF4a, Foxa1/2, Zhx2, HNF6, Mtf1, SEK1, Nfkb1, c-Jun, Itih-4, and Hex. To answer this question exactly, however, too few abstracts mention particular Theiler stages or days post congestion. They rather refer to “early stages of development”, and the exact time span might be presented in the full text article only.

Reconstructing outcomes of large-scale screening

Thut et al. provided a list of 62 genes found expressed during eye development in mice, together with developmental stage and substructure (Thut et al., 2001). Of the 62 genes, 26 were not previously reported (as of 2001); to 16 genes, novel valuable information could be added; 20 genes were fully reported before. Expression patterns were summarised for E12.5, E13.5, E14.5, E16.5, E18.5 and P2. Using MousePubMed, we tried to reconstruct the result of this large-scale screen of 1000 genes.

Complete EMAP test set

To evaluate capabilities of automated searches against the complete EMAGE data, the experimental setting was as follows. Genes in EMAGE have annotated tissues, in which they were detected at various stages of embryo development. Thus, we queried MousePubMed with each gene and checked which tissues were mentioned

³ Important for answering this query are returned PubMedIDs 12736215, 12055180, 11403717.

⁴ Important are PubMedIDs 16540514, 16412590, 14978045, 14684990, 14516667, 12889070, 9879712, 9115712.

Type of information	Amount of data
Genes with tissues, stages	1437
Genes with at least one non-trivial tissue, stages	1346
Triples of gene, tissue, stage	18,179
Triples of gene, non-trivial tissue, stage	12,782
Tuples of gene, non-trivial tissue	8653

Table 8.4. Overview on data sets contained in EMAGE. EMAGE contains associations of genes and tissues to developmental stages.

Type of information	Amount of data
Triples of gene, non-trivial tissue, stage	1637 (12.8%)
Tuples of gene, non-trivial tissue	2667 (30.8%)
Genes with at least one tissue and stage	537 (37.4%)

Table 8.5. Quantification for facts on gene/tissue/developmental stages retrieved from literature. Number of tuples/triples consisting of gene and tissue or gene, tissue and stage found in PubMed abstracts retrieved by the query “mouse AND development.”

in the resulting PubMed abstracts. This was based on co-occurrence of the gene considering, a tissue, and a Theiler stage (day) in the same abstract. Currently, there are 1437 genes in the EMAGE database annotated with (sometimes multiple) tissues and stages. All in all, we identified 18,179 such triples — gene, tissue, and stage — in EMAGE. Many of the annotations consist of general annotations for tissue, like “mouse”, “embryo”, “left”, “female”, “node”. We removed such trivial instances, because they were very frequently found. 12,782 triples referred to specific tissues, and we tried to find these triples using the aforementioned term extraction (also see Table 8.4).

8.3.3 Results

Ontologies are widely used for annotation. They are also useful for literature search, but the extraction of terms from text is a difficult problem due to the complexity of natural language.

As Table 8.5 shows, we were able to reconstruct 31% of the gene-tissue associations in EMAGE using PubMed abstracts. Only 13% of the full information (gene, tissue, exact stage) was contained in abstracts. All in all, the data recovered from PubMed included information on about 37% of the EMAGE genes. We noted that in many cases, abstracts do not mention specific time points during development. Sometimes, “early” and “late development” are mentioned, which we resolved as described previously in this section. On the other hand, mentions like “in early liver development” could not be resolved to specific overall-stages without background information. Cross-checks revealed that indeed much of the necessary information was only mentioned in the full text of references annotated by EMAP for a specific association. These figures are encouraging as only abstracts are used (Wächter et al., 2007).

8.4 LMOPubMed

LMOPubMed is a web-based search engine that allows searches PubMed and categorises documents using terms from the newly created Lipoprotein Metabolism Ontology (LMO). The goal of LMOPubMed was to categorise lipids with respect to risk factors for diseases and ethnic specifics.

Reference:

Alexopoulou, D.*, Wächter, T.*, Pickersgill, L., Eyre, C., and Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*, 9(Suppl 9):S2, Impact factor 2009: 3.7 *shared first author

The creation of the Lipoprotein Metabolism Ontology and LMOPubMed

The Lipoprotein Metabolism Ontology (LMO) was manually built in collaboration with domain experts from Unilever for the purpose of document retrieval. It consists of 522 concepts and 964 additional synonyms, with an average term length of 15 (2 words of 7,5 characters).

There have been two challenges specifically during the creation of LMOPubMed, namely the amount of synonyms and syntactic variations which are in this domain very important to be recognised, and the consistent modelling of human ethnic groups.

Lipoprotein subclasses based on particle size, buoyant density, composition, etc. are specifically difficult to differentiate, as there do not exist clear limits between them. Depending on the way of measurement and the difference in surface lipid content, they can be expressed in different ways. For example, in the case of LDL, there are 5 different subclasses based on particle size (LDL I-V), but there are also references such as “small dense LDL” or “buoyant LDL” that are very often found in text but could contain a mixture of different subclasses. Since we need to keep only a simple hierarchy with parent-child relationships, we do not incorporate any “definitional” information (e.g. that “small dense LDL” consists of a mixture of LDLIII and LDLIV). In these cases, we put the synonyms according to the authors’ use, for example “small dense LDL” as a synonym for LDL III and “buoyant LDL” or “large LDL” as synonyms for LDL I. Additionally the handling of term variation had to be regarded. Terms like “LDL I”, “LDL-I”, “LDL-1”, “LDL 1”, “LDL1” and “LDLI” are also variants of the same term. The process of manually inserting such lexical variants (with hyphens, apostrophes, slashes, or even American/British spelling variants) in the ontology is tedious and time-consuming. Automatic learning methods did help significantly.

Ethnic groups may result in inconsistencies with implication on reasoning as described in the following example: a researcher is interested in the different lipoprotein levels in patients of different race and geographical location, since there has been evidence that these two factors affect lipoprotein metabolism. Combination of geographical information as well as racial information in one part of the ontology is, therefore, needed. Many articles refer to “African-Americans” as “blacks”,

The screenshot displays the LMOPubMed web application. On the left, the 'Induced LMO Ontology' is shown as a hierarchical tree. The root node is 'diabetes type 2', which branches into 'LMO Ontology [13]', 'Phenotype [13]', 'Diabetes [13]', 'Type 2 [5]', 'Body composition [5]', 'Type 1 [2]', 'Metabolic syndrome [1]', 'Hypertriglyceridaemia [1]', 'Coronary Heart disease [1]', and 'Hypotriglyceridaemia [1]'. Further sub-nodes include 'Cell [6]', 'Human [6]', 'Liver [1]', 'Protein [6]', 'Mass [2]', 'Expression [2]', 'Activation [1]', 'Secretion [1]', 'Transcription factors [1]', 'Affinity [1]', 'State [5]', 'fasted [5]', 'fed [1]', 'Physicochemical properties [4]', 'Composition [3]', 'Secretion [1]', 'Animal model [1]', 'ex vivo [1]', 'in vivo [1]', 'Subclass [2]', '(Partic) distribution [1]', 'Concentration [1]', 'Genetics [2]', 'Lifestyle [2]', and 'Diet [2]'. A 'Find Term' button is located at the top of this ontology.

The main content area on the right shows search results for the query 'diabetes type 2' and LMO term 'LMO Ontology'. It includes a search bar with the query, a 'Search' button, and a 'more options' link. Below the search bar, it lists 'Frequent terms for query "diabetes type 2" are: Diabetes(13) Lean(5) Type 2(5) fasted(5) Diet(2)'. The results are divided into two sections: 'LMO terms' and 'Highlighted keywords: diabetes, type, 2'. The 'LMO terms' section lists 8 terms: Diabetes (100%), Cell (100%), adult (100%), Genetics (100%), Expression (100%), Transcription factors (100%), Lean (100%), and Type 2 (99%). The 'Highlighted keywords' section lists 3 terms: Lean (100%), Diabetes (100%), and male (100%).

The search results also include a list of articles found for the query. The first article is titled 'Transcriptional control of pancreatic endocrine cell development.' and is by Gasca R. The second article is titled '[Glycemic control in prepubertal and pubertal patients with diabetes type 1 - a one year ambulatory follow-up]' and is by Gomes MB, Castro SH, Garfinkel T, Fernandes LM, Cunha EF, Lobbio VI.

Fig. 8.8. Screenshot of LMOPubMed. Like in GoPubMed, the left side shows the LMO Ontology with documents automatically assigned to each node. The right side shows the documents itself with the LMO terms highlighted.

so the term must be included under “ethnic group”. Then the following must be valid: define “Caucasian”, “African” and “Asian” as “ethnic group”, “American” is a “Caucasian”, “African-American” is a “African”, “African-American” is a “American”, “African-American” is “black” (synonym), “Caucasian” is white (synonym) but “African-American” cannot be “Caucasian” or “white” (although he is “American”). This is similar to the case of mammals that lay eggs or the “Man, Woman, Eunu- ch”; people very often formulate rules such as “normally is-a”, as there are always exceptions. For the LMO we excluded the “American” concept and added “African- American” as child of “African” and “Hispanic-American” as child of “Caucasian”.

8.5 Go3R

Go3R is a web-based search engine that actively retrieves 3Rs relevant information. It provides the search result automatically linked to an "intelligent table of contents". Thereby, Go3R actively supports the user in finding information on alternative methods that is available on the internet.

Reference:

Sauer, U. G. *, Wächter, T. *, Grune, B., Doms, A., Alvers, M. R., Spielmann, H., and Schroeder, M. (2009). Go3R - semantic internet search engine for alternative methods to animal testing. *ALTEX*, 26(1):17–31, *Impact factor 2009: 1.3* *shared first author

8.5.1 Introduction

The role of information retrieval for the application of the 3Rs principle

The search for alternative methods to animal testing is legally mandatory, but also economically advisable. In 2008, nearly 2.7 million vertebrate animals were used for scientific purposes in Germany (BMELV, 2008). The numbers are expected to increase dramatically with the new EU Chemicals Regulation *REACH* (EC 1907/2006). *REACH* stands for **R**egistration, **E**valuation, **A**uthorisation and **R**estriction of Chemicals. Once all regulations contained in *REACH* are in place, all companies manufacturing or importing chemical substances into the European Union in quantities of one tonne or more per year will be required to register these substances with a new European Chemicals Agency in Helsinki, Finland. For this registration the manufactures have to provide detailed information on the chemicals' potential impacts on both human health and the environment which will lead to an increase in animal testing. Many substances that have already been used for a long time will have to be registered and possibly tested by 2018. It is expected that *REACH* will lead to an EU-wide increase in the number of animals used of up to 400,000 animals per year (Höfer et al., 2004). Estimations by Hartung and Rovida (2009) are even 10 times higher with in total of 54 million vertebrate animals used within the next 11 years to test 68,000 substances. The costs of testing have been estimated at up to 5.4 billion Euro (Fauser, 2007).

REACH

EU Directive 86/609/EEC for the protection of laboratory animals (Commission of the European Communities, 1986) obliges scientists to consider whether any planned animal experiment can be substituted by other scientifically satisfactory methods not entailing the use of animals or entailing less animals or less animal suffering, before performing the experiment. The ongoing revision of EU Directive 86/609/EEC is expected to lead to even more stringent rules regarding the evaluation of the indispensability of animal experiments in the course of licensing procedures (TEWG, 2003; Commission of the European Communities, 2008).

Thus, the Replacement, Reduction, and Refinement of animal experiments in accordance with the *3Rs principle* (Russell and Burch, 1959) is a mandatory obligation - morally, legally, and also economically. To meet this obligation, scientists must consult the relevant scientific literature in respect to potential alternative methods prior to conducting any experimental study using laboratory animals.

3RS PRINCIPLE

my search

- Show Clipboard [0]
- Find Specific Concept
- Previous Queries
- Current Query
- AND
- 3Rs Animal Use Alternatives [1,384]
- 3Rs Relevant (AUTOMATIC CLASSIFICATION)

what

- 3Rs Categories
 - 3Rs Related Categories [1,384]
 - 3Rs Replacement Alternative Methods [192]
 - Pharmacology, Replacement Methods [109]
 - Food Hygiene, Replacement Methods [50]
 - BBB Pharmacology - animal use replacement
 - Pharmacy, Replacement Methods [30]
 - Physiology, Replacement Methods [13]
 - Immunology, Replacement Methods [13]
 - Microbiology, Replacement Methods [6]
 - 3Rs Refinement Alternative Methods [13]
 - Pharmacy, Refinement Methods [5]
 - Physiology, Refinement Methods [5]
 - Immunology, Refinement Methods [5]
 - Food Hygiene, Refinement Methods [3]
 - 3Rs Reduction Alternative Methods [21]
 - Acute Oral Toxicity Reduction Method [15]
 - Pharmacy, Reduction Methods [5]
 - Immunology, Reduction Methods [5]
 - Food Hygiene, Reduction Methods [1]
 - 3Rs Relevant (AUTOMATIC CLASSIFICATION)
- Knowledge Base

who

- All Authors [1,368]
 - Baskettter DA [51]
 - Kimber I [42]
 - Dearman RJ [24]
 - Kimber I [21]
 - Roberts DW [19]
 - Gerberick GF [16]
 - Cronin MT [17]
 - Combes R [17]
 - Grindon C [16]
 - Spielmann H [15]
 - Garrod JF [14]
 - Baskettter DA [13]
 - Scholes EW [12]
 - Gerberick GF [12]
 - Ryan CA [11]
 - Robinson MK [11]
 - Spielmann H [10]
 - Dearman RJ [9]
 - Liebsch M [9]
 - Cosmetic Ingredient Review Expert Panel [9]
 - Hilton J [8]
 - Kandárová H [8]
 - Liebsch M [8]
 - Klopman G [7]
 - Kern PS [7]
 - more
- Find specific author...

where

- Earth [1,384]
- All Journals [1,384]
 - Other Journals [1,380]
 - High Impact Journals [4]
 - Find specific journal...
 - Reviews only

when

- All Times [1,384]
- Publication date
 - Last Day [0]
 - Last Week [1]
 - Last Month [16]
 - Last Year [105]
 - Last 5 Years [480]

Home GoGene GoPubMed GoWeb Ontology Editor Help

"3Rs Animal Use Alternatives"[go3r] "3Rs Relevant (AUTOMATIC CLASSIFICATION)"[go3r] **find it**

1,384 documents semantically analyzed

statistics

documents

In vitro study of silica nanoparticle-induced cytotoxicity based on real-time cell electronic sensing system.
 Yang, Hong, et al.
 Journal: *J Nanosci Nanotechnol*, Vol. 10 (1): 561-8, 2010
 The above experimental results were compared with the experiments of a tetrazolium compound-based colorimetric method (MTT assay) and LDH activity in medium.
 Affiliation: School of Public Health, Southeast University, Nanjing 210096, PR China.

Xenobiotic Metabolism Gene Expression in the EpiDerm(TM) In Vitro 3D Human Epidermis Model Compared to Human Skin.
 Hu, T. et al.
 Journal: *Toxicol In Vitro*, 2010
 We report comparison of the expression of 139 genes encoding xenobiotic metabolizing enzymes in the EpiDerm(TM) model and human skin.
 Affiliation: The Procter & Gamble Company, Miami Valley Innovation Center, P.O. Box 538707, Cincinnati, OH 45253 (M. Aardema currently Marilyn Aardema Consulting, LLC, 5315 Oakbrook Dr Fairfield OH 45014).

Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro.
 Emter, Roger, et al.
 Journal: *Toxicol Appl Pharmacol*, 2010
 In vitro tests are needed to replace animal tests to screen for the skin sensitization potential of chemicals.
 Affiliation: Givaudan Schweiz AG, Ueberlandstrasse 138, CH-8600 Duebendorf, Switzerland.

Comparison of different cytotoxicity measurements for the in vitro micronucleus assay using L5178Y and TK6 cells in support of OECD draft Test Guideline 487.
 Lorge, Elisabeth
 Journal: *Mutat Res*, 2010
 The reference genotoxic agents mitomycin C, cadmium chloride, 2-aminoanthracene, vinblastine sulphate and 5-fluorouracil were tested in the in vitro micronucleus assay, in mouse lymphoma L5178Y cells and in human lymphoblastoid cells TK6, without cytokinesis block.
 Affiliation: Servier Group, Drug Safety Assessment, Gidy, France.

An evaluation of a cultured human corneal epithelial tissue model for the determination of the ocular irritation potential of pharmaceutical process materials.
 Seaman, Christopher W, et al.
 Journal: *Toxicol In Vitro*, 2010
 These results infer that HCE cultures, alone or as a part of a tiered hazard screening programme, have promise for use in reducing reliance on live subject tests and contribute to generation of an appropriate hazard classification and label advice.
 Affiliation: GlaxoSmithKline, Park Rd., Ware, Hertfordshire, UK SG12 0DP.

BALB/c 3T3 cell transformation assay for the prediction of carcinogenic potential of chemicals and environmental mixtures.
 Mascolo, Maria Grazia, et al.
 Journal: *Toxicol In Vitro*, 2010
 An in vitro cell transformation assay (CTA) is proposed as an alternative to in vivo carcinogenicity testing.
 Affiliation: Environmental Carcinogenesis and Risk Assessment, Environmental Protection and Health Prevention Agency - Emilia-Romagna Region (ER-EPA), Viale Filopanti 22, 40126 Bologna, Italy.

Fig. 8.9. Screenshot of Go3R. The left side shows the intelligent table of contents with categories of alternative methods, authors, locations, journals, and publication dates.

Consideration and incorporation of all available scientific information is a crucial part in the planning of any scientific project. As regards the question of whether or not to perform an animal experiment in the course of a planned biomedical research project, it is not only scientific standard, but also a legal requirement to base this decision on the best available information. However this scientific standard and legal obligation can only be met, if all those involved in the planning, licensing and performance of biomedical research are able to obtain all available relevant information on alternative methods in accordance to the 3Rs principle. This intricacy demonstrates how closely the request to replace, reduce, and refine animal experiments is connected to information retrieval and, as a result, information technology.

The core of any scientific strategy or political incentive to refine, reduce and replace animal experiments lies in the availability of relevant information regarding alternative methods.

In a feasibility study funded by the National German Centre for Documentation and Evaluation of Alternatives to Animal Experiments (ZEBET) at the Federal Institute for Risk Assessment (BfR) in Berlin, Transinsight GmbH, Dresden, in co-operation with the Biotechnology Center of the Technical University Dresden, ZEBET at the BfR, Berlin, and Scientific Consultancy - Animal Welfare, Neubiberg/Munich, set out to develop Go3R, a 3Rs knowledge-based internet search engine, and to evaluate whether this new semantic technology tool could serve to improve internet inquiries on alternative methods. As a result of the feasibility study, a prototype of the search engine Go3R has been made available online via <http://Go3R.org> free of charge in April 2008.

State of the art and current technological problems in retrieving information on alternative methods from the internet

Currently, the procedure of determining the availability or unavailability of alternative methods – as required by law whenever a scientist plans to perform an animal experiment – is complex and the different steps taken by the scientist in pursuing this task are oftentimes not transparent to others. Millions of potentially relevant documents are scattered across the internet, patent databases, literature databases, and intranets. Classical search technologies seek exactly what is asked for without using the context of the search terms or other information which might be relevant for the search query. Therefore they are unable to reveal alternatives that the user has not explicitly searched for. Finally, there are no methodologies to ensure that scientists, animal welfare officers, and authorities truly base their decisions on all available relevant information. Ontology-based literature provides efficient and comprehensive access to all information stored as text. In the following, these problems are to be explained in further detail.

- **Range of biomedical information available in the internet**

A large variety of databases and websites with information on alternative methods exists (Hakkinen and Green, 2002). This includes classical databases, such as AGRICOLA, AGRIS, BIOSIS Previews, CAB Abstracts, EMBASE and PubMed/MEDLINE. Furthermore, specialised added-value alternatives databases, such as AnimAlt-ZEBET and ECVAM/DB-ALM, and an abundance of websites with information on animal experiments and alternatives or measures in line

with the 3Rs principle are available, such as the websites of DG Environment of the EU Commission, European Chemicals Bureau, Council of Europe, European Food Safety Authority, Animal Welfare Information Center, AltTox, Fund for the Replacement of Animals in Medical Experiments, UK National Centre for the Replacement, Refinement and Reduction of Animals in Research, Swiss Federal Agency for Veterinary Affairs, German National Ministry for Education and Research, German Federal Ministry for Nutrition, Agriculture and Consumer Protection, Netherlands Centre for Alternatives to Animal Use, Netherlands National Institute for Public Health and Environment, Federation of European Laboratory Animal Science Associations .

The listed websites contain very diverse and unequally processed information – and oftentimes only concerning certain aspects of the 3Rs concept – and mostly have to be queried and scanned one-by-one by the searching scientist.

- **Internet searches in the area of alternative methods**

It is the inherent challenge of any search to find those documents in the vast pool of different information resources that are relevant for replacing, reducing and refining planned animal experiments. This problem becomes even more evident, when looking at the concrete amounts of data available on the internet. For instance, PubMed/MEDLINE encompasses 18,590,000 documents (as of 01/2009); and EMBASE encompasses 12,773,576 documents (as of 01/2009). Both databases are updated every day. Even though the contents of different literature databases overlap, they are far from being identical, because different journals are indexed by different databases. Furthermore, the indexing of publications and other information on alternative methods is limited by the following situation: During indexing, mostly just the keywords of a given publication are being used, while the question if the methodologies depicted in the publication might be relevant for replacing or reducing animal experiments is not taken into consideration⁵. This implies that articles which do not explicitly mention that they present an alternative method will not be indexed as animal use alternatives (Grune et al., 2004).

A fundamental problem regarding searches is the selection of those terms that are to be searched for in the data pool. If the terms selected are too general, the numbers of documents retrieved will be far too large; and the documents are not sufficiently relevant. If the terms are too specific, important documents will be excluded from the list of results.

8.5.2 Development of the Go3R ontology

The initial step for the creation of Go3R was the creation of the 3Rs domain ontology. A first frame for this ontology was created by identifying existing ontologies and vocabularies relevant to the topic which were the AGRICOLA thesaurus and the branches diseases, anatomical structures and organs or chemical compounds from the Medical Subject Headings (MeSH). With regard to the planned comprehensive

⁵ Nelson, S. J., The Alternative Project. <http://www.nlm.nih.gov/mesh/presentations/publicr/ppframe.htm>

navigation structure, it was considered necessary and meaningful to include both specifically 3Rs relevant terms in the ontology (e.g. “Local Lymph Node Assay”) as well as thematic-defining terms (e.g. “dermatitis, allergic contact”). Parts of existing ontologies relating to 3Rs relevant terms on the housing and handling of laboratory animals as well as to physiological and psychological conditions of laboratory animals and to certain 3Rs methods (AGRICOLA) and to thematic-defining terms (MeSH) were linked to the Go3R Ontology. Descriptions of the existing ontologies modifications are given in Table 8.6. The ontology was further extended with newly composed 3Rs relevant terms. For this purpose, own expert knowledge and vocabulary from the documents in the ZEBET database AnimAlt were used. As a preliminary framework requiring trial in practice, 28 different branches were defined and created for the Go3R ontology. It was distinguished between thematic-defining and directly 3Rs relevant branches (Table 8.6):

Ontology branch	Definition of terms listed in branch	Examples for terms included in the branch	Comment
3Rs Institutions	Institutions with the primary mission to make a contribution towards replacing, reducing, or refining animal tests and animal experiments.	e.g. “ZEBET”, “ECVAM”	Terms are directly 3Rs relevant.
3Rs Methods in the Life Sciences	Concrete 3Rs test methods sorted in accordance to their area of use in the life sciences.	e.g. the Neutral Red Uptake Phototoxicity Test is listed under “3Rs in toxicology - 3Rs in photoirritation” and the HPLC method for Calcitonin determination under “3Rs in pharmacy”	Terms are directly 3Rs relevant.
3Rs Relevant	Special term in the ontology without child-terms	No child-terms due to special status of this term	Filter with which the search engine grades the 3Rs relevance of the documents retrieved during a query
3Rs Research Projects	Names of research projects pursuing the primary goal to develop 3Rs test methods	e.g. ReProTect for the EU integrated project aiming at developing “a novel approach in hazard and risk assessment of reproductive toxicity”, www.reprotect.eu	Terms are directly 3Rs relevant.
Animal Care and Handling	Procedures with which humans care for animals or handle and manipulate them	e.g. “group housing”, “ad libitum feeding”, “animal identification” or “capturing of animals”	Word sense disambiguation required for each individual term.
Animal Conditions, Physiological or Psychological	Desirable or undesirable physiological or psychological states of animals	e.g. “animal behaviour”, “animal welfare”, “animal distress”	Word sense disambiguation required for each individual term.
Animal Experiments	Animal models, animal test methods and names of in vivo refinement or reduction methods	e.g. “disease models, animal”, “guinea-pig maximisation test”, “fixed dose procedure”	Thematic-defining branch.

Ontology branch	Definition of terms listed in branch	Examples for terms included in the branch	Comment
Animal Species	Species of invertebrate and vertebrate animals. In the case of vertebrate animals, emphasis is given to mentioning those species that are regularly used in experiments and for other scientific procedures.	e.g. "rabbit", "rodent", "non-human primate"	Word sense disambiguation is required to instruct the search engine to distinguish whether reference to animal species in a document means that the respective animals were used in vivo or that e.g. primary cells of such animals were used in vitro.
Animal Use Alternatives	Classification of animal use alternatives in accordance to their correlation to the 3Rs principle.	"Reduction alternative", "refinement alternative" and "replacement"	In contrast to the branch "3Rs methods in the life sciences", in which concrete test methods are listed in accordance to their areas of use, the branch "animal use alternatives" merely maps the fundamental distinction between "replacement methods", "reduction methods" and "refinement methods".
Bioethics	Terms relating to the 3Rs principle as such as well as to other bioethical topics as the case may be.	"3Rs principle"	
Biological Material & Organisms for Animal Use Alternatives	Cell, tissues, organs and single-cell organisms employed in non-animal test methods.	e.g. cultivated primary cells or specific cell lines, organ and tissue cultures, reconstituted organs	Documents in which reference is made to specific cell lines, for instance, have a strong likelihood of containing 3Rs relevant information.
Body Systems & Structures	Anatomical systems and structures.	e.g. "gastrointestinal tract", "blood vessels" & MeSH branch "Body Systems"	Structures" and additional terms, e.g. relevant terms relating to veterinary medicine that were not contained in MeSH.
Diseases & Symptoms	Disorders of structure of function of the human or animal body.	e.g. "endocarditis", "leukaemia"	MeSH branch "Diseases & Symptoms" with some addition terms where needed.
In Vitro Culture Technology & Equipment	Concepts relating amongst others to (1) in vitro cell culture systems, (2) cell culture additives, (3) cell culture equipment or (4) manipulations with cells.	e.g. (1) "suspension culture", (2) "serum free medium", (3) "perfusion systems", (4) "cell cryoconservation"	In the respective context, such terms refer to 3Rs relevant information with a high probability.
In Vitro Experimental Design	Terms describing the experimental design of in vitro test methods, including (1) test endpoints, (2) endpoint detection methods and (3) cell culture test scoring procedures.	e.g. (1) "cell viability", "DNA damage", "enzyme induction", (2) "neutral red uptake" and (3) "half maximal inhibitory concentration, IC50".	In a given context, such terms can refer to 3Rs relevant information.

Ontology branch	Definition of terms listed in branch	Examples for terms included in the branch	Comment
In Vivo Experimental Design	Terms describing the experimental design of in vivo test methods.	e.g. the dosage of animals, or the scoring of test results	This branch requires further elaboration to put emphasis on terms that make reference to the application of humane endpoints or to measures towards reducing the numbers of animals used in procedures.
Laboratory Science	Animal The science and technology dealing with the procurement, breeding, care, health, and selection of animals used in biomedical research and testing.	-	This branch requires further development to include relevant terms relating to the science dealing with the care and use of animals used in biomedical research and testing.
Laboratory Animals	Specific types of laboratory animals.	e.g. "specific pathogen-free animals"	Further experience in practice is required to establish its usefulness in practice and to adapt it accordingly.
Life Sciences	Terms describing the sciences concerned with the study of living organisms.	e.g. "food hygiene", "microbiology", "toxicology"	This is a thematic-defining branch from MeSH mapping concepts on life science terms relevant for the issue of 3Rs methods.
Method Specification	Attributes describing types of methods.	e.g. "in vitro" and "in vivo"	Many documents describing cell culture test methods, for instance, include the term "in vitro" so that the narrowing down of search results to this term might enable a first broad selection of possibly relevant articles.
Methodology	Terms and concepts for specific test methodologies	e.g. "enzyme-linked immunoassay", "high performance liquid chromatography"	This branch is conceived to supplement the branch "3Rs in the Life Sciences" aiming to enable documents describing specific methodologies to be retrieved independently from their application of use.
Product Properties & Effects	Characteristics of products and their wanted or unwanted effects	e.g. "liver toxicity", "biocompatibility"	
Product Testing & Assessment	Terms describing certain forms of testing and assessment of products	e.g. "efficacy testing", "risk assessment"	

Ontology branch	Definition of terms listed in branch	Examples for terms included in the branch	Comment
Statistics	Terms describing the science of collecting, summarizing, and analyzing data that are subject to random variation.	e.g. “predictive value”, “uncertainty factor”	This branch, currently including relevant on statistics from MeSH, requires further elaboration to include specific terms relating to the statistical evaluation of test results that are specifically relevant for the 3Rs ontology.
Substances, Preparations & Products	Terms referring to both biological substances and industrially produced substances, preparations and products	Such terms enable sub-sorting search results in accordance to specific biological substances under investigation (e.g. specific transmitters, enzymes, etc., evaluated for instance in biomedical studies) or in accordance to the type of test substance (heterocyclic compounds, polycyclic compounds, etc.) used in toxicological studies.	This is a thematic-defining branch, imported from MeSH.
Toxic Actions of Substances	Concepts and terms that describe substances in regard to their category of harmful action on living organisms	e.g. “irritant”, “mutagene”	This is a predominantly thematic-defining branch, imported from MeSH.
Toxicity Testing Strategies, 3Rs	Terms describing testing strategies making a contribution to refining, reducing or replacing animal testing as such.	e.g. “tiered testing strategy”, “integrated testing strategy”	When such terms are used in a document, the information is very likely to be 3Rs relevant.
Validation of Test Methods	Terms describing the different steps and aspects of validation of test methods.	e.g. “reproducibility, test methods”, “predictivity, negative, test methods”	In combination with further specific terms, such terms can point to 3Rs relevant documents.

Table 8.6: List of the 28 branches of the Go3R ontology prototype.

In order to sort the newly composed and conferred terms, the 3Rs relevant and the thematic-defining vocabulary was grouped into concepts and a hierarchy was defined. While forming the respective branches of the ontology, strict attention was paid to correctly adhering to the respective necessary subdividing steps and to labelling and defining the terms as precisely as possible so that correct correlations and mappings to superordinate terms could be achieved. In order to obtain a strictly hierarchical “parent-child” relationship between terms, all child terms and sub-child terms of a given branch of the ontology not only have to be children of their respective immediate superordinate term, but at the same time also sub-children of all higher direct superordinate terms of the respective higher terms of the given branch. The assignment of superordinate and subordinate concepts revealed the necessity to

Found	Test method	in AnimAlt
✓	luminescent bacteria toxicity test	ZEBET40
✓	red blood cell (RBC) test	ZEBET 30
✓	chorioallantoic membrane vascular assay (CAMVA)	ZEBET 272
✓	EYTEX	ZEBET 271
✓	fluorescence leakage test	ZEBET 270
✓	neutral red release (NRR) assay	ZEBET 265
✓	neutral red uptake (NRU) cytotoxicity assay	ZEBET 26
✓	chicken egg chorioallantoic membrane (HET-CAM) assay	ZEBET 25
	silicon microphysiometer	ZEBET 245
✓	human skin cell multilayer cultures	ZEBET 237
✓	low volume eye test (LVET)	ZEBET 236
	optical function of bovine lens	ZEBET 109
✓	chicken enucleated eye test (CEET)	ZEBET 107
✓	isolated rabbit eye (IRE) test	ZEBET 105
✓	bovine corneal opacity and permeability (BCOP) assay	ZEBET 103
	pollen tube growth test	ZEBET 101

Table 8.7. Listing of sixteen 3Rs methods to determine eye irritation effects of substances existing in the ZEBET database.

create a further type of correlation between terms in the ontology in addition to the assignment of direct parent-child relationships. Thereby, an article in which e.g. a concrete 3Rs method is not explicitly mentioned could still be recognised as relevant for the specific topic searched for in an indirect manner, for example if it mentions specific cells, endpoints or endpoint detection methods which would be relevant for the respective application.

Example 8.1 (Eye irritation). The very general search query eye irritation was used to search for publications on methods with which the in vivo Draize eye irritation test on the rabbit eye might be replaced, reduced or refined, as the case may be. The search term was deliberately chosen to lack further terms narrowing down the search query so that a comprehensive picture of the capability and effectiveness of the search engine itself in narrowing down the search result in respect to 3Rs methods and methodologies could be obtained. The search query eye irritation was chosen as a first detailed use scenario since the ZEBET database AnimAlt lists a large variety of 3Rs methods for the determination of eye irritating effects of substances and provides an abundance of literature references on this topic. The information provided for in the ZEBET database was taken as “reference data” to test whether the ontology would enable retrieval of the same amount of information. There exist sixteen 3Rs methods to determine eye irritating effects of substances (Table 8.7).

Figures 8.10 and 8.11 illustrate the different steps of performing and evaluating the search query eye irritation. The 651 documents retrieved as a result of the search query are presented together with the intelligent table of contents. By clicking through this table of contents, the user can extract sub-lists relating to a topic that is of interest to him (e.g. the 67 articles relating to “3Rs in Toxicology”, Figure 8.10). Within this sub-list, a further sub-list of 12 articles relates to the “BCOP test” 8.11. In the resulting “intelligent table of contents”, 39 of these 651 documents were listed in the ontology branch “3Rs Methods in the Life Sciences” under the term “3Rs in Eye



Fig. 8.10. Go3R user interface with the search query “eye irritation” indicated in the search field. The screenshot presents the sub-list of 67 articles relating to “3Rs in Toxicology” extracted from the full search retrieval of 651 citations found for the query “eye irritation” (as of 05/2010). The nodes in the “intelligent table of contents” are sorted by the number of assigned documents. The documents on the right side are sorted that sentences which contain query terms or selected nodes are ranked higher. In the first document this is “eye irritation” and “BCOP assay”, a descendant of “3Rs in Toxicology” (Figure 8.11).

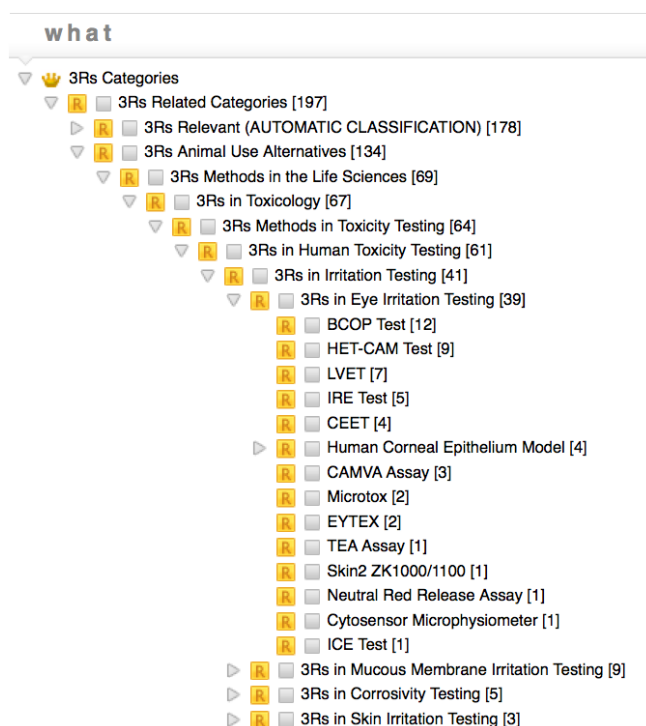


Fig. 8.11. Refinement in Go3R using the intelligent table of contents. Within the sub-list of 67 articles related to “3Rs in Toxicology”, a further sub-list of 12 articles relates to the “BCOP test” (as of 05/2010).

Irritation Testing”. 37 of these 39 documents indeed contained information on 3Rs methods to determine eye irritating effects. They provided information on the following 12 of the 16 3Rs methods listed in the ZEBET database: *NRU test*, *IRE test*, *HET-CAM*, *BCOP*, *NRR assay*, *human epithelial cell line tests*, *fluorescence leakage test*, *luminescent bacteria toxicity test*, *LVET*, *EYTEX*, *CAMVA*, *CEET* (Table 8.7). Additionally, documents were retrieved on *TOPKAT* and *EpiOcular*, for which there are no separate entries in the ZEBET database. Information on a further 3Rs eye irritation test, the *RBC test*, was found in the ontology branch “*Non-Animal Laboratory In Vitro Bioassay*” (containing a sub-list of 58 retrievals). Thus, the Go3R ontology allowed targeted retrieval of information on 13 of the 16 3Rs eye irritation methods listed in the ZEBET database (Table 8.7).

Coverage At the same time, the search query showed that PubMed only covers approximately one third of the literature on 3Rs methods to determine eye irritating effects provided for in the ZEBET database. One reason for this discrepancy is that PubMed has only lately taken up indexing a number of 3Rs relevant journals (for instance, *ALTEX* and *Toxicology In Vitro* have been indexed as of the year 2000 and *ATLA* as of Jan/Feb 2001.) Other journals, such as *In Vitro Toxicology* (which was published from fall 1986/1987 until winter 1997) are not indexed by PubMed at all. As was mentioned above, two publications of the 31 documents of the search retrieval on “*Eye Irritation*” which Go3R listed under “*3Rs in Toxicology*” did not contain information on 3Rs eye irritation test methods. These two documents provided information on 3Rs methods for other endpoints instead: the “*Local Lymph Node As-*

say” to determine sensitising effects and the “*basal cytotoxicity test*” to determine acute systemic toxicity. Nevertheless, the abstracts of both publications contained the term “*eye irritation*”. As a result the search engine “correctly” retrieved these two publications during the search query eye irritation and listed them under the term “*3Rs in Toxicology*”. In conclusion, the Go3R search engine prototype was unable to retrieve information on three of the 16 3Rs methods listed in the ZEBET database with the search query eye irritation. To identify the reason for this, in a next step these three methods were explicitly searched for with the specific search queries silicon microphysiometer, optical function of bovine lens test, and pollen tube growth test. These specific search queries revealed that the respective publications presenting these three methods (as far as they were listed in PubMed at all) referred to the determination of eye irritating effects with terms and concepts that had not yet been included in the ontology, such as “ocular safety testing”, “irritancy screening”, “irritating potential of ingredients of cosmetic formulations”. These terms were included into the ontology in order to become able to retrieve such information via Go3R as well.

Example 8.2 (“Blood-Brain-Barrier”). The search query for term “*Blood-Brain-Barrier*” was chosen as another example to test the Go3R ontology and search engine in retrieving relevant 3Rs information in the area of fundamental biomedical research. Again, the search term “*Blood-Brain Barrier*” was deliberately chosen to be very general lacking further terms narrowing down the search query to ensure that only the effectiveness of the Go3R search engine was tested. In the classical search

The screenshot displays the Go3R user interface. On the left, a sidebar titled "my search" shows a hierarchical tree of "3Rs Categories". The "what" section is expanded, showing a list of categories with counts. A red arrow points from the "In Vitro Blood-Brain-Barrier Methods [312]" category to the search results. Below this are sections for "who", "where", and "when". The main content area shows the search results for the query "blood brain barrier 'In Vitro Blood-Brain-Barrier Methods'[g]". It indicates that 312 of 26,361 documents were semantically analyzed. Below this, there are two document entries. The first entry is titled "Prostanoid EP(1) receptor antagonist reduces blood-brain barrier leakage after cerebral ischemia." and lists the author Fukumoto, Ken-ichi, et al. The second entry is titled "Magnetic nanoformulation of azidothymidine 5'-triphosphate for targeted delivery across the blood-brain barrier." and lists the author Saied, Zainulabedin M, et al. Both entries include the journal name, PMID, and a brief abstract.

Fig. 8.12. Go3R user interface, with the search query “Blood-Brain Barrier” indicated in the search field. Within the respective sub-list, 312 articles related to “In Vitro Blood-Barrier Methods”.

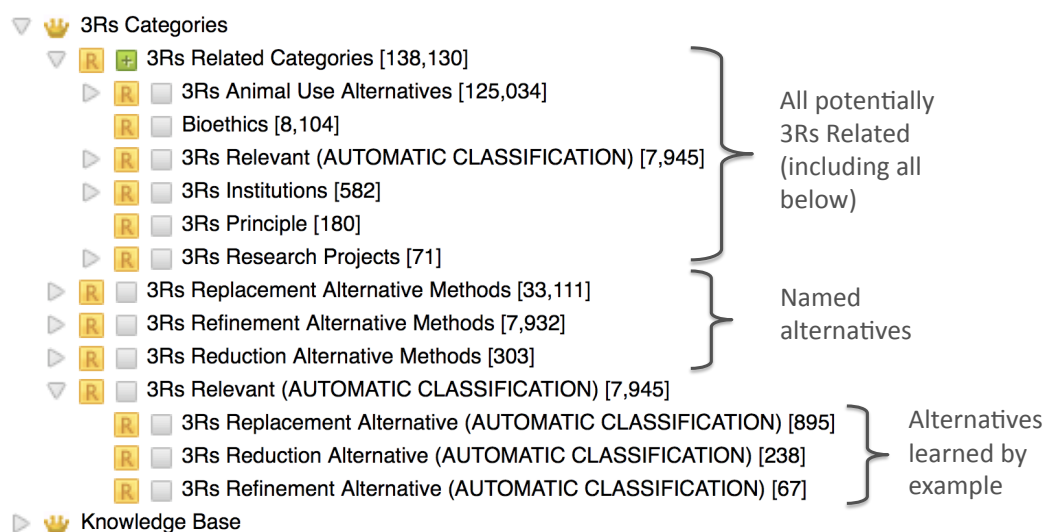


Fig. 8.13. 3Rs Related Categories in Go3R. The top section under WHAT in the dynamic table of contents lists categories for named alternative methods and the automatically classified methods which have been learned by example.

in PubMed, the search query Blood-Brain Barrier resulted in a total of 26,361 documents (as of 05/2010), which, evidently, are presented in the form of a long list. Additional search steps are required to narrow down the number of hits in the search of 3Rs relevant documents. The same search query performed with Go3R (Figure 8.12) also resulted in 26,361 documents, however the accompanying “intelligent table of contents” allowed to quickly extract 312 articles relating to “*In Vitro Blood-Brain Barrier Method*” from within this vast data pool.

8.5.3 3Rs relevance filter

In the ontology, the term “3Rs Relevant” has been assigned a special status as categorising term for all documents which show relevance to the domain of animal testing alternative methods. Thus the term “3Rs Relevant” serves as a filter.

3Rs relevance signet 3Rs relevant articles are labeled with the “3Rs relevance signet” assigned by the 3Rs relevance filter (Figures 8.10). The green signet indicated that the domain expert has labeled this document as relevant. The grey signet indicates the with a maximum of three stars the classification confidence of the machine learning approach.

Machine learning

Technically, the 3Rs relevance filter categorises documents using a machine learning technology called Maximum Entropy Method (Berger et al., 1996). The Maximum Entropy Method enables the filter to learn the characteristics of documents belonging to a certain pre-defined category. Provided a large amount of training examples, the algorithm automatically extracts a set of relationships inherent in the examples,

and then combines these rules into a model of the data that is both accurate and compact. Relationships can be found based on textual occurrences of terms as well as meta information on documents such as the publishing journal or the year of publication. The learned model is used to assign each unknown document to the pre-defined category.

In regard to the 3Rs relevance filter, the Maximum Entropy Method was trained using 2,346 PubMed abstracts hand-annotated as being 3Rs relevant and 2,346 abstracts half randomly selected from all PubMed documents and half randomly selected from the journals hand-annotated as being not 3Rs relevant, thereby teaching the search engine to distinguish between 3Rs relevant and 3Rs irrelevant documents based on the words and document meta-information contained and associated with the documents. As a result, the 3Rs filter, currently available, highlights those documents in which methods or methodologies are depicted that can make a contribution towards replacing, reducing and refining animal experiments. The performance of the classification was evaluated using a cross validation with test and training sets and empirically against self defined queries to retrieve 3Rs relevant queries.

Cross validation is a technique for assessing whether the results of a statistical analysis can hold as generalisation for a bigger independent data set. The results allow to compare different parameter settings of the tested algorithm. In several rounds subsets of documents are randomly selected as training and test set. The training set is used to train the machine learning model which is then tested to classify the test set in matching an non-matching (positive and negative) documents. For the performed 5-fold cross validation the positive and negative document set were splitted in ten parts. In five rounds alternating four parts were used for training and one part for testing. This 5-fold cross validation showed for the 3Rs relevance filter a F-measure of 0.91. With 0.93 precision was higher than recall of 0.86. It can be expected that most classifications are correct, but also a number of 3Rs relevant articles will be not marked as such. The decision to favour precision over recall has been made to create initial user acceptance for the search engine.

Run	Samples	Recall	Precision	F-measure
1	940	0.86	0.97	0.91
2	938	0.84	0.96	0.90
3	938	0.87	0.95	0.90
4	938	0.90	0.96	0.92
5	938	0.86	0.95	0.90

Table 8.8. 5-fold cross validation for the classification of “3Rs Relevant” documents using Maximum Entropy model classification. The validation for 2,346 positive curated documents and 2,346 negative curated documents lead to precision of 0.95 – 0.96, recall of 0.84 – 0.90, and a F-measure of 0.91 – 0.92. The threshold for the classification was 0.1. Results for thresholds 0.001 and 0.2 are listed in Table 10.12.

Practical evaluation of the 3Rs relevance filter

To show that 3Rs relevant documents are labeled as such, a number of sample queries have been evaluated with Go3R. The queries have been chosen to retrieve documents that are likely to be 3Rs relevant. It has been checked how many of the retrieved documents have been automatically classified as 3Rs relevant.

Hypothesis: All documents retrieved for the PubMed query for the synonyms of 3Rs Reduction, Refinement, and Replacement alternative method are 3Rs relevant. (search in title or abstract)

```
"3Rs principle"[tiab] OR "3Rs concept"[tiab]) OR ("replace"[tiab]
AND "reduce"[tiab] AND "refine"[tiab]) OR ("replacement"[tiab] AND
"reduction"[tiab] AND "refinement"[tiab]) OR ("Russell"[tiab] AND
"Burch"[tiab]))
```

201 of 207 documents categorised as 3Rs relevant (97%)

Of the 232 documents analysed 201 were categorised as 3Rs relevant. A manual check of the 31 missed documents revealed that 24 were truly not 3Rs relevant (Table 10.11).

Hypothesis: All documents for MeSH term "*Animal Testing Alternatives*" in MEDLINE are 3Rs relevant:

```
Animal Testing Alternatives[MESH]
```

1,893 of 1,909 documents categorised as 3Rs relevant (99.1%)

1,574 documents annotated with "*Animal Testing Alternatives*" till 2008 have been used for training the machine learning classifier.

Hypothesis: All documents for MeSH term "*Animal Testing Alternatives*" in MEDLINE are 3Rs relevant (with abstracts only):

```
Animal Testing Alternatives[MESH] hasabstract[text]
```

1,230 of 1,254 documents categorised as 3Rs relevant (98.1%)

1,574 documents annotated with "*Animal Testing Alternatives*" till 2008 have been used for training the machine learning classifier.

Hypothesis: All documents annotated with Go3R term "*Embryonic Stem Cell Test*" in MEDLINE are 3Rs relevant. The embryonic stem cell test represents a validated alternative method for in vivo embryotoxicity testing.

```
"Embryonic Stem Cell Test"[go3r]
```

44 of 50 documents analysed are 3Rs relevant (80%)

44 documents were correctly annotated as 3Rs relevant. 3 documents are 3Rs relevant, but were not classified as such. The other document PMID:14734052 mentioned "*embryonic stem cell test system*" but not the explicit test.

Hypothesis: All documents annotated with Go3R term "*BALB/c 3T3 NRU Assay*" in MEDLINE are 3Rs relevant. The 3t3 NRU assay is an alternative method for the assessment of phototoxic hazard of cosmetic products.

```
"BALB/c 3T3 NRU Assay"[go3r]
```

16 of 22 documents categorised as 3Rs relevant (73%)

Hypothesis: All documents with affiliation containing the National German Centre for Documentation and Evaluation of Alternatives to Animal Experiments (ZEBET) at the Federal Institute for Risk Assessment (BfR) in Berlin are 3Rs relevant (with abstracts only):

```
zebet[ad] hasabstract[text]
```

44 of 46 documents categorised as 3Rs relevant (96%)

The missing documents were not primarily 3Rs relevant:

PMID: 19157061 *“On the impact of the molecule structure in chemical carcinogenesis.”*

PMID: 20358685 *“High-molecular weight protein toxins of marine invertebrates and their elaborate modes of action.”*

Hypothesis: All documents annotated written by Horst Spielmann, a senior author in the field, are 3Rs relevant (with abstracts only):

```
Spielmann H[au] Berlin[ad] hasabstract[text]
```

46 of 57 documents categorised as 3Rs relevant (80.1%)

None of the 11 documents not annotated as 3Rs relevant mentions 3Rs related terms of the Go3R ontology.

For the majority of tested potentially 3Rs relevant document sets a high percentage ($\geq 94\%$) of correct classification was observed. For the specific test methods *“Embryonic Stem Cell Test”* (80%) and *“BALB/c 3T3 NRU Assay”* (73%) a number of documents from foundational research have not been classified as 3Rs relevant. While this still has to be investigated it does not have a negative influence on the overall retrieval, as the terms are modelled as named alternative methods and will as such be regarded in searches for 3Rs related information.

8.5.4 Recognition of Go3R ontology terms in text

The current version of Go3R uses like GoPubMed an word alignment algorithm (Doms, 2004). The text and the terms are decomposed into token stems. A local sequence alignment algorithm (Smith and Waterman, 1981) is used to map term tokens to text tokens. Penalty values for gaps, deletions and insertions were experimentally calculated. The word alignment is adjusted by the information value for a word. This value is in GoPubMed the based on the frequency of the occurrence of words in the ontology. For Go3R corpus frequencies from PubMed have been used.

The algorithm was tested on 100 manually curated MEDLINE abstracts and achieved good results (89.5% precision and 81.4% recall). The algorithm has a quadratic runtime, the nature of all approaches based on dynamic programming. Processing of a MEDLINE abstract took in 2004 about 10ms upon fresh annotation. 10.000 new articles new 100 second for annotation. This algorithm does not disambiguate the meaning of words. Therefore false annotations of short ambiguous ontology concepts occur in the current Go3R, unless a disambiguation model was trained on examples as explained below.

8.5.5 Disambiguation for 3Rs methods

The 3Rs relevance filter classifies the general relevance of documents for the domain. The Maximum Entropy Method (*MaxEnt*) (Berger et al., 1996) was used for the disambiguation of 3Rs terms. First the candidate documents are collected. All documents containing any of the synonyms of a term are candidates, e.g. any of 15 for “*3Rs Reduction Alternative*” in Table 8.11. Second, the *MaxEnt* classifier is applied and documents are either accepted as belonging to a term or are dismissed.

Currently, *MaxEnt* models have been trained for the general method types “*3Rs Replacement Method*”, “*3Rs Reduction Alternative*”, and “*3Rs Refinement Method*”, as well as for the specific terms “*3Rs Principle*” and “*CASE*”. Definitions for the senses used in Go3R have been listed in Table 8.10.

Disambiguation of specific terms

Results of the 5-fold cross validation for the terms (Table 8.10; compare also Section 8.5.3 (Cross validation)) shows that specific terms like “*CASE*” and “*3Rs Principle*” can be found with high precision and good recall. For *CASE* a F-measure > 0.96 was obtained. *CASE* (the QSAR method) can be perfectly separated from e.g. *CASE REPORT*. Positive hints of the *MaxEnt* model were e.g. Computer, Evaluation, activity, compound, predict. Negative hints were report, disease, examination, therapy.

For 3Rs principle results are slightly lower. Still a precision of 0.86 – 1.00 and a recall of 0.64 – 0.84 have been achieved in the different runs.

Disambiguation of Replacement, Reduction, and Refinement

While the results for the disambiguation for specific terms are very good. The results for the disambiguation of the higher level 3Rs methods differ significantly between each other. The best results in terms of F-measure have been achieved for “*3Rs Replacement Alternative Methods*” (0.83-0.93), moderate for “*3Rs Reduction Alternative Methods*” (0.58 – 0.77), and low for “*3Rs Refinement Alternative Methods*” (0.40 – 0.73). This corresponds to the difficulty of finding these terms for which definitions are listed in Table 8.9. Replacement is the simplest task as Go3R currently regards every in vitro method described with specific named cell lines as a potential replacement method. Reduction is more difficult, especially with respect to precision, because documents that mention “reduction of animals” can be equally likely replacement alternatives, and also to some extent refinement alternative methods. Especially Replacement and Reduction are hard to distinguish. To make the matter worse, training documents have been collected in a way that positive documents for one of the three Rs were used as negative documents for the other two Rs drawing the fine line between the methods and making cross validation difficult.

Ontology term based disambiguation

When collecting curations for “*3Rs Reduction Alternative Methods*” it has been found that positive curations were annotated with a number of 3Rs related terms, while negative curations were not. The hypothesis is that the disambiguation of reduction, refinement, and replacement can be performed much better when using the

Term	Definition
3Rs Principle	The 3Rs principle was introduced by Russel and Burch in 1959 who motivated the (R)eplacement, (R)efinement, and (R)eduction of animal testing in their book “The principles of humane experimental techniques”. The 3Rs principle in Go3R means all mentions of the principle or references to all the three Rs without the need to name or describe a specific method.
3Rs Replacement Method	Replacement means the substitution for conscious living higher animals of insentient material.
3Rs Reduction Method	Reducing the number of animals used to obtain information of a given amount and precision, or increasing the amount of useful data obtained from the same number of animals, without compromising the quality or the quantity of animal-based research. Three main ways for reducing animal use: a) better research strategy; b) better control of variation; c) better statistical analysis.
3Rs Refinement Method	Refinement means any decrease in the severity of inhumane procedures applied to those animals which still have to be used.
CASE	CASE stands for Computer Automated Structure Evaluation and is a method for quantitative structure-activity relationships prediction (QSAR).

Table 8.9. Definitions for selected disambiguated terms in Go3R. The definitions for the terms Replacement, Reduction, and Refinement by Russell and Burch (1959) will be further refined in the ongoing Go3R project.

ontology terms as features for the *MaxEnt* machine learning classifier. To evaluate this for the term “*3Rs Reduction Method*” the 50 latest user curations from all 152 negative curation made for initially misclassified documents have been reviewed for document-wise 3Rs related co-occurring annotations. Same was done for the 50 latest user curations of the 78 positively curated documents. Documents and terms are listed for the negative curations in Table 8.13 and for the positive curation in Table 8.12. Without any statistics, it is clearly visible that positively curated documents have not only many more terms annotated but also many more 3Rs related terms annotated. Negatively curated document for “*3Rs Reduction Alternative Methods*” rarely have a 3Rs related terms annotated. Hence, an extension of the disambiguation approach to consider ontology terms found in title and abstract is likely to be highly advantageous for the overall disambiguation performance.

(a) Replacement					(b) Reduction				
Run	Samples	Recall	Precision	F-measure	Run	Samples	Recall	Precision	F-measure
1	202	0.77	1.00	0.87	1	32	0.81	0.72	0.77
2	202	0.71	1.00	0.83	2	32	0.75	0.80	0.77
3	202	0.87	1.00	0.93	3	32	0.44	0.88	0.58
4	202	0.76	0.99	0.86	4	30	0.53	0.89	0.67
5	200	0.79	1.00	0.88	5	30	0.73	0.73	0.73

(c) Refinement					(d) 3Rs principle				
Run	Samples	Recall	Precision	F-measure	Run	Samples	Recall	Precision	F-measure
1	28	0.64	0.69	0.67	1	50	0.84	0.88	0.86
2	28	0.43	0.86	0.57	2	50	0.64	0.94	0.76
3	28	0.50	0.88	0.64	3	50	0.76	0.95	0.84
4	28	0.28	0.67	0.40	4	48	0.83	1.00	0.91
5	26	0.62	0.89	0.73	5	48	0.79	1.00	0.88

(e) CASE				
Run	Samples	Recall	Precision	F-measure
1	26	1.00	0.93	0.96
2	24	1.00	1.00	1.00
3	24	1.00	1.00	1.00
4	24	1.00	0.92	0.96
5	24	1.00	1.00	1.00

Table 8.10. Cross validation results for the disambiguation of terms in Go3R. Specific terms like CASE (mean F-measure = 0.98) and 3Rs principle (mean F-measure = 0.85) can be disambiguated. From the 3Rs methods, replacement (mean F-measure = 0.87) can be better disambiguated than Reduction (mean F-measure = 0.70) or refinement (mean F-measure = 0.60).

Synonyms for 3Rs Reduction Alternative Method	
animal testing reduction	reduction of the number of animals
less animals	reduces number of animals
reduction of animal testing	reduction, animal testing
reduction test method	reduction alternative
animal use reduction	reduced number of animals
reducing animal testing	reduce animal models
reduction of animal use	reduce animal use
fewer animals	

Table 8.11. Listing of the synonyms of the term “3Rs Reduction Alternative”. All synonyms found in the manually curated documents have been added to the term in Go3R. Documents containing any of the synonyms are disambiguated based on positive and negative training examples. Broader synonyms like “fewer animals” have been added to increase the document prior disambiguation.

PMID	all terms	3Rs terms	List of 3Rs terms
11797832	7	1	Test Method
11846632	10	4	Animal Welfare; Sensitisation; Skin Sensitisation; Test Method;
11890466	20		Regulatory Acceptance; Safety Assessment; Topical Application
12449363	8		Magnetic Resonance Imaging; Positron Emission Tomography;
12734637	15	7	ATC method; Acute toxicity; Biometrics; Inhalation Toxicity; Test Method; Stepwise procedure; Toxicity
12843228	14	4	Anesthesia; Number of Animals; Stress; Acute Toxic Class Method
15026810	13	0	
15053500	22	7	Anesthesia; Anesthetics; Animal Handling; Animal Numbers; Stress; Toxicity; Toxicology
15057405	33	6	Cell Lines; Cytotoxicity; in vitro; Non-Animal In Vitro Bioassay; Safety Testing; Toxicology;
15057409	20	2	Animal Numbers; Caco2 Cells; Cell Culture; Cell Model; Cells, Cultured; In vitro; Non-Animal In Vitro Bioassay; Primary Cells; Primary Culture; Test Method
15519906	10	2	Linear regression Models; Toxicology
15570743	13	1	Screening; Test Method;
15601222	6	1	Animal Numbers
15601231	18	9	3Rs Principle; Acute Oral Toxicity; Acute Toxicity; Acute Toxic Class Method; Animal Numbers; Safety Testing; Suffering; Test Method; Toxicity
15703127	3	1	Animal Numbers
15719147	13	5	3Rs Principle; Animal Numbers; Distress, Animals; Narcosis; Test Method
15896439	13	9	Acute Oral Toxicity; ATC method; Acute Toxicity; Animal Numbers; Biometrics; Oral Toxicity; Safety Testing; Test Method; Toxicity
16426021	2	1	Animal Numbers
16708692	13	6	Risk Assessment; Systemic Toxicity; Test Method; Tiered Testing Strategy; Toxicity; Toxicology
16708695	16	7	Adverse Effect; Risk Assessment; Safety Testing; Systemic Toxicity; Test Method; Toxicity; Toxicokinetics
16885064	8	1	Animal Numbers
16945419	15	3	Anesthesia; Anesthetics; Animal Numbers
16988468	13	6	Animal Numbers; Bioluminescence; Fluorescence; in vitro, MRI; PET
17088988	30	10	A549 Cells; Animal Numbers; Animal Welfare; Cell Lines; Comet Assay; HT29 Cells; HT29 Cells; Lung Cell Lines; Lung Cells; Test Method
17454397	18	3	Biocompatibility; Magnetic Resonance Imaging; Positron Emission Tomography
17500484	8	0	
17559313	7	4	Animal Numbers; Animal Welfare; Integrated Testing Strategy; Test Method
17645410	5	2	Animal Numbers; Human-Animal Relations
18304838	22	1	Test Method
18360728	11	5	Cell Culture; Cells, Cultured; ex vivo; in vitro; Non-Animal In Vitro Bioassay
18370307	12	1	Animal Numbers
18522474	4	1	Test Method
18551236	12	4	Computer Methodology; Functional Magnetic Resonance Imaging; Magnetic Resonance Imaging; Positron Emission Tomography
18606234	16	2	Test Method; Toxicity
18758243	20	5	Animal Numbers; Non-Animal In Vitro Bioassay; Sensitisation; Test Method; Toxicity
18931182	15	2	Linear Regression Models; Test Method
19025337	18	11	3Rs Animal Use Alternatives; Animal Numbers; Chronic Toxicity; Decision-Tree Testing Strategy; FRAME; Integrated Testing Strategy; Non-Animal Laboratory Test Methods; Repeated-Dose Toxicity; Safety Testing; Test Method; Toxicity
19237454	12	0	
19292572	4	3	Animal Numbers; Animal Welfare; FRAME
19292573	2	1	FRAME
19292574	3	2	Animal Numbers; Laboratory Animals
19379807	11	2	Developmental Toxicity; Toxicity
19409482	13	3	Immunotoxicity; Neurotoxicity; Toxicokinetics
19426798	7	4	Developmental Toxicity; One-Generation Study; Safety Testing; Test Method
19432769	11	7	Adverse Effect; Animal Numbers; Institutional Animal Care and Use Committees; Potency Testing; Safety Testing; Test Method; Toxicity
19540332	5	2	Toxicity; N3RC
19651214	10	3	Anesthesia; Test Method; Tumor Volume (Tumor Burden)
19665509	9	5	Animal Experiments; Animal Numbers; Animal Welfare; Risk Assessment; Safety Testing
19738022	8	2	Animal Model; Animal Numbers; Toxicology; Aerosol
19765670	16	2	ECVAM Workshop; Regulatory Acceptance; Toxicology; Positron Emission Tomography;

Table 8.12. Fifty most recent positively curated documents for “3Rs Reduction Alternative”. The majority of document contain more than one term from the Go3R Ontology.

PMID	all terms	3Rs terms	List of 3Rs terms
19829501	1	0	
19829505	0	0	
19864737	12	1	Test Method
19864740	0	0	
19864742	2	0	
19864757	14	0	
19864759	3	0	
19885300	22	1	Test Method
20011037	3	1	Computer Simulation
20028390	2	0	
20041845	11	2	transgenic mice; Anesthetic Effect
20051206	4	1	Test
20053282	10	0	
20056586	21	1	Toxicity
20060281	2	0	
20065518	4	1	Mass Spectrometry; LC-MS
20068695	0	0	
20069583	0	0	
20069735	0	0	
20077018	6	1	MCF-7 Cells
20082346	4	0	
20087208	7	3	Computer Simulation; Electroencephalogram; Test Method
20092660	6	0	
20107611	5	0	
20120391	3	0	
20131764	2	0	
20132601	10	0	
20135079	3	1	Test Method
20138670	18	5	ELISA; MTT Assay; Assay; in vivo; in vitro
20144349	3	1	Study Duration
20145060	10	0	
20156146	4	0	
20160097	16	1	Magnetic Resonance Imaging
20161583	5	0	
20168451	2	0	
20170749	12	1	Test Method
20171248	3	0	
20171418	9	0	
20172800	4	0	
20174122	5	0	
20183679	1	0	
20184771	2	1	Test
20185011	9	0	
20191771	1	1	Magnetic Resonance Imaging
20198955	5	0	
20208407	4	0	
20217589	4	1	Positron Emission Tomography
20358267	3	0	
20394393	3	0	
20397214	4	0	

Table 8.13. Fifty most recent negatively curated documents for “3Rs Reduction method”. In comparison to Table 8.12 only very few terms from the Go3R ontology have been found in the documents. The next versions of Go3R use ontology terms as features for the machine classification.

8.5.6 Autor curation

Training data is needed for both, disambiguation of 3Rs terms and classification using the 3Rs relevance filter. In Go3R this training data is collected with the Curation Tool (Figure 8.14). In total positive or negative curations have been collected for 12,520 documents (Table 8.14).

Term	Number of curations
3Rs Relevant	9133
Replacement	1029
Refinement	561
3Rs Principle	495
CASE	244
TIMES	158
	147

Table 8.14. Overview on the number of manual curations for terms in Go3R. The table shows the terms in Go3R with > 10 curations.

[In silico, in vitro, in omic experimental models and drug safety evaluation]

PMID: 19154704 Related Articles

Claude, Nancy, Goldfain-Blanc, Françoise, Guillouzo, André

Journal: Med Sci (Paris), Vol. 25 (1): 105-10, 2009

Over the last few decades, **toxicology** has benefited from scientific, technical, and bioinformatic developments relating to **patient safety assessment** during clinical and **drug** marketing studies. Based on this knowledge, new **in silico**, **in vitro**, and "omic" experimental models are emerging. Although these models cannot currently **replace** classic safety evaluations performed on **laboratory animals**, they allow compounds with unacceptable **toxicity** to be rejected in the early stages of **drug** development, thereby **reducing the number of laboratory animals** needed. In addition, because these models are particularly **adapted to mechanistic studies**, they can help to improve the relevance of the data obtained, thus enabling better prevention and **screening** of the **adverse effects** that may occur in **humans**. Much progress remains to be done, especially in the field of validation. Nevertheless, current **efforts** by industrial, academic laboratories, and regulatory agencies should, in coming years, significantly improve preclinical **drug safety** evaluation thanks to the integration of these new methods into the **drug** research and development process.

Affiliation: Institut de Recherches Internationales Servier, 6, place des Pléiades, 92415 Courbevoie, France. nancy.claude@fr.netgrs.com

Pubmed MeSH: Biological Markers, Cell Differentiation, Genome, Metabolism, Proteome, Quantitative Structure-Activity Relationship, Transcription, Genetic

Wikipedia: 3 R's, Adverse effect, Concise, Delton, Drug safety, Drug toxicity, Drugs, Effort, Exertion, Human, In-vitro, Laboratory Animals, Patients, Screening, Three Rs, Toxic effects, Toxicity, Toxicity testing, Toxicology

3Rs Principle
3Rs Reduction Alternative (AUTOMATIC CLASSIFICATION)
3Rs Relevant (AUTOMATIC CLASSIFICATION)
Adverse Effect
Number of animals (Animal Numbers)
Adaptic (Bisphenol A-Glycidyl Methacrylate)
Drug Safety (Drug Toxicity)
Drugs
Effort (Exertion)
Human
in silico
in vitro
Laboratory Animals
Mechanistic Study (Mechanistic Test)
Patients
Safety Assessment (Safety Testing)
Screening
Toxicity
Toxicology
replacement alt **search**
3Rs Replacement Alternative (AUTOMATIC CLASSIFICATION)

Fig. 8.14. On-line Curation Tool in Go3R. All manual and automatic term assignments are shown next to a document. The icons +, -, o are used to give feedback. "o" stands for neutral and is submitted together with a comment created using the commenting option. In the screenshot a negative curation has been made for term "3Rs Reduction Alternative", because the phrase "reducing the number of laboratory animals" has been wrongly classified. Instead a positive curation has been added for term "3Rs Replacement Alternative" as an in vitro replacement method is described.

8.5.7 Term generation for Go3R

The term generation in DOG4DAG can provide abbreviations and lexical variants required to locate Go3R terms in text.

Abbreviations One example where DOG4DAG finds abbreviations is “*botulinum toxin type A*”. From PubMed abstracts the has been found ‘BoNT-A’, ‘BTXA’, ‘BTA’, ‘BoNT/A’, ‘BTX-A’, and ‘BoNTA’. More examples of abbreviations for 3Rs related terms found with DOG4DAG are listed in Table 8.15. In cases where the long form of a terms occurs only once, while the short form occurs with high frequency, the short form is preferred as term label, e.g. “Hen’s Egg Test-Chorioallantoic Membrane” is generated as synonym to “HET-CAM”.

Lexical variants Including the abbreviations, 20 different lexical variants can be found for the term “*botulinum toxin type A*”, namely ‘botulinum toxin types A’, ‘botulinum toxin-A’, ‘botulinum toxin type A’, ‘botulinum neurotoxin serotypes A’, ‘Botulinum Toxin Type A’, ‘Botulinum toxin A’, ‘botulinum neurotoxin A’, ‘BoNTA’, ‘Botulinum toxin type A’, ‘botulinum neurotoxin a’, ‘BoNT-A’, ‘botulinum toxin A’, ‘Botulinum toxin-A’, ‘Botulinum Toxin-A’, ‘Botulinum Toxin A’, ‘Botulinum Toxin Type-A’, ‘Botulinum neurotoxin A’, ‘BoNT/A’, ‘Botulinum neurotoxin serotype A’, and ‘Botulinum Neurotoxin A’. More examples of lexical variants found with DOG4DAG are listed in Table 8.16. The term generation method aims to retrieve long terms, e.g. “Joint Research Centre-Institute for Health and Consumer Protection”, and preserve and not truncate chemical formulas e.g. “(3RS)-nerolidyl diphosphate”.

8.5.8 Definition extraction for Go3R

DOG4DAG can provided high quality definitions for many terms. To show this we tried to semi-automatically defined all 152 method terms that exist below “*3Rs methods in Toxicity Testing*”. For 65 (43%) terms good definitions were found. Table 8.17 shows that for 27% of terms the top retrieved definition was suitable. Five example definitions are listed below. The generated definitions for all terms can be found in OBO format in Table 8.17.

- **Local lymph node assay** (1st retrieved definition) – *Local lymph node assay is an animal-based toxicology test developed as an alternative to the transdermal guinea pig sensitization test.*
- **EYTEX** (1st retrieved definition) – *Eytex is an alternative testing method that evaluates eye irritancy of a protein alteration system by using an in vitro, or test tube, procedure.*
- **In Vitro Skin Absorption Test** (1st retrieved definition) – *In Vitro Skin Absorption Test is a full replacement for the in vivo skin penetration test under OECD TG 428.*
- **Embryonic Stem Cell Test** (3rd retrieved definition) – *embryonic stem cell test is an in vitro screening assay used to investigate the embryotoxic potential of chemicals by determining their ability to inhibit differentiation of embryonic stem cells into spontaneously contracting cardiomyocytes.*
- **EpiDermTM** (1st retrieved definition) – *EpiDermTM is a commercially available human skin model consisting of normal human-derived epidermal keratinocytes (NHEK), which have been cultured to form a multilayered, highly differentiated model of the human epidermis.*

Term	Lexical variants
botulinum toxin type A	'BoNT-A', 'BTXA', 'BTA', 'BoNT/A', 'BTX-A', 'BoNTA'
enzyme-linked immunosorbent assay	'ELISA'
C-terminal fragment of <i>Clostridium perfringens</i> enterotoxin	'C-CPE'
diarrhetic shellfish poisoning	'PSP'
embryonic stem cell test	'EST'
skin integrity function test	'SIFT'
quantitative structure-activity relationships	'QSARs', 'QSAR'
low volume eye test	'LVET'
fish embryo toxicity test	'FET'
Food and Drug Administration	'FDA'

Table 8.15. Abbreviations extracted for animal testing alternatives related terminology.

Term	Lexical variants
EpiSkin	'episkin', 'EPISKIN', 'Episkin', 'EpiSkin'
botulinum toxin type A	'botulinum toxin types A', 'botulinum toxin-A', 'botulinum toxin type A', 'botulinum neurotoxin serotypes A', 'Botulinum Toxin Type A', 'Botulinum toxin A', 'botulinum neurotoxin A', 'BoNTA', 'Botulinum toxin type A', 'botulinum neurotoxin a', 'BoNT-A', 'botulinum toxin A', 'Botulinum toxin-A', 'Botulinum Toxin-A', 'Botulinum Toxin A', 'Botulinum Toxin Type-A', 'Botulinum neurotoxin A', 'BoNT/A', 'Botulinum neurotoxin serotype A', 'Botulinum Neurotoxin A'
prevalidation	'pre-validation', 'prevalidation', 'Prevalidation'
ecotoxicity	'eco-toxicity', 'ECOTOXICITY', 'Ecotoxicity', 'ecotoxicity'
IC(50)	'IC(50)', 'inhibitory concentration 50%', 'IC(50)s'
HET-CAM	'Hen's Egg Test-Chorioallantoic Membrane', 'hen's egg test-chorioallantoic membrane', 'HET-CAM'
microRNA	'microRNAs', 'micro-RNAs', 'micro ribonucleic acid', 'microRNA', 'Micro-RNA', 'micro RNA', 'MicroRNAs', 'miRNAs', 'MiRNAs'
(3RS)-nerolidyl diphosphate	'(3RS)-nerolidyl diphosphate'
murine embryonic stem cells	'Murine embryonic stem cells', 'murine ES cells', 'murine embryonic stem cells', 'mESC'

Table 8.16. Lexical variants extracted for animal testing alternatives related terminology.

Good definition retrieved within rank	Number of terms	Percent of all terms
1	41	27%
2	55	36%
5	58	38%
all	65	43%

Table 8.17. Number of terms out of 152 terms in branch 3Rs in Human Toxicity Testing of the Go3R ontology for which a definition could be semi-automatically created. For 65 (43%) of the human toxicity testing methods a definition could be found, for 41(27%) the top ranked generated definition has been chosen.

8.5.9 Summary and Discussion

In April 2008, Go3R has been made available online free of charge under <http://www.Go3R.org>. It aims to enable all those involved in the planning, authorisation and performance of animal experiments to determine the existence of non-animal methodologies in a fast, comprehensive, and transparent manner. Go3R is the first and currently only semantic tool with a specific focus on alternative methods. Recently, other semantic search technologies have been developed and made available online which also mine the tremendous pool of biomedical information in the internet. Nevertheless, the search benefit achieved by Go3R in retrieving information on alternative methods in accordance to the 3Rs principle cannot be paralleled by any other of the currently available semantic search engines. The most important difference highlighting the uniqueness of Go3R in searching for alternatives to animal experiments is its expert knowledge-based 3Rs specific ontology which specifically maps subjects and terms related to animal use alternatives. Thereby, retrieved information is classified with a focus on alternative methods in a meaningful manner. In contrast, the ontology-based search engine GoPubMed (see above) covers general biomedical issues using the Gene Ontology and the Medical Subject Headings (MeSH). Therefore, it does not serve to specifically retrieve 3Rs methods.

Go3R's contribution Go3R aims to optimise the practice of determining either the availability or non-availability of 3Rs methods in all scientific areas in which animals are being used, except for education. It is expected that the efficient utilisation of alternative methods documented in the scientific literature will result in a quantifiable reduction of the numbers of animals used in scientific procedures. In order to substantiate this expectation and to be able to provide concrete figures regarding a reduction of laboratory animals as a result of improved information retrieval technologies, the further development of Go3R shall include concrete use scenarios determining the number of laboratory animals saved by each individual scenario due to the improved methodology for information retrieval. Thereby, evidence shall be provided that and how many laboratory animals are saved because of the improved information retrieval system in individual cases. Regarding the avoidance of animal tests for regulatory purposes, e.g. to meet the requirements of the EU Chemicals Regulation (Commission of the European Communities, 2006) or those of the EU legislation on plant protection products (Commission of the European Communities, 1991), researchers both have to search for existing data in order to avoid repetitive animal testing and for relevant 3Rs methodologies. The importance of thorough internet searches for existing data on toxicological endpoints in preventing animal testing has been confirmed during the United States High Production Volume Chemicals Programme (Nicholson et al., 2004). Go3R can make a significant and unique contribution to finding both types of information. With the help of its detailed ontology branches on chemical substances, information on existing data can be extracted and sorted in a fast and transparent manner, while the information on available alternative methods can be extracted and sorted with the respective ontology branches in which the 3Rs relevant knowledge is mapped.

3Rs Relevance Filter and disambiguation The 3Rs Relevance Filter is the first step to categorise documents for their relevance in the area of alternative methods to

animal experimentation. The cross validation results with precision > 0.95 and recall > 0.84 are promising and the recall obtained for selected example queries supports these promising technical results with already 9 out of 10 relevant documents that have been automatically categorised as 3Rs relevant.

The disambiguation method used in Go3R is very well capable to decide on the correct sense of the two specific terms CASE and 3Rs principle where models have been trained so far. To some extent the method is capable to capture the specifics of the of the 3Rs, Replacement, Reduction, and Refinement. Especially difficult is the separation of Replacement and Reduction as well as the correct determination of Refinement. It has been shown per example for the term “*3Rs Reduction Alternative Methods*”, that an extension of the attributes regarded by the machine learning method to ontology terms found in documents, will be highly beneficial for the classification accuracy.

Semi-automated ontology generation It has been shown on example from the domain, that the DOG4DAG ontology generation methods developed in this thesis are applicable in the domain of animal testing alternative. They can support the ontology engineer with relevant terms, relevant abbreviations, and lexical variants. The definition generation method can find definitions for half of the 152 evaluated alternative methods for 3Rs toxicity testing.

8.5.10 Future work

With the start of the BMBF funded research project “Entwicklung und Etablierung einer Semantischen Suchmaschine für Alternativmethoden zu Tierversuchen (FK:614 40003 0315489A)” in continuation of this thesis’ work, further development can be devoted to www.Go3R.org.

3Rs method classification In further work the specific classification of 3R methods as Reduction, Replacement, or Refinement method as well as a specification of methods as “in vitro” or “in vivo” method will be approached.

Re-ranking by relevance Currently, Go3R labels documents as relevant to 3R. Further work will show whether query dependent re-ranking of retrieved documents by 3Rs relevance will be beneficial.

Transparent document retrieval The Go3R search engine already indicates why documents are retrieved by highlighting the direct or indirect associations made using the ontological model. Further work will improve this by adding reports on the indications which lead to 3Rs relevance classification, either by exploiting linguistic relations between the ontology terms or references to external facts provided by reliable sources such as the ZEBET database.

Richer ontology model The current Go3R ontology supports only one type of subsumption relationship. The experience building Go3R has shown, that richer semantic modelling, e.g. new relationship types or distinct types of synonymy, are desirable and possibly will help to improve document classification.

Sources for documents Currently Go3R searches PubMed which is only a part of the information on alternative methods available. Other specialised databases and web resources need to be included to obtain a comprehensive overview on information regarding alternatives to planned animal experiments.

Community platform The greater goal is to create a community platform, where scientist can point to alternatives, highlight their publication, link information and search for colleagues working on similar fields. This will include:

- collaborative methodology reviews
- collaborative ontology editing
- collaborative ontology curation
- collaborative rule building

8.6 Contributions in the development of semantic search applications

All work on the applications was shared work. Many applications have been developed. In the following my specific contributions will be listed:

Go3R

Design and implementation of several Go3R prototypes to specifically target the search on information on alternatives to animal experiments. This includes the work on the Go3R ontology supporting Dr. Ursula Sauer, the specification of the 3Rs relevance filter method extending work of Dr. Andreas Doms. I was responsible for the project management of the feasibility study funded by the National German Centre for Documentation and Evaluation of Alternatives to Animal Experiments (ZEBET) at the German Federal Institute for Risk Assessment (BfR) in Berlin and Transinsight GmbH, Dresden. The study financed the creation of the www.Go3R.org semantic search engine prototype and lead to the extension of the funding provided by the BfR. Additionally the work will be continued in cooperation with BASF SE Ludwigshafen, in the ongoing BMBF funded project “Entwicklung und Etablierung einer Semantischen Suchmaschine für Alternativmethoden zu Tierversuchen (FK:614 40003 0315489A)” .

Go3R Ontology Editor

Idea and specification of the user interaction and GUI layout for an easy to use editor for simple ontologies to create the background knowledge for Go3R. The first editor software has been implemented by Hick’n’Hack Software GbR⁶, the second version of the editor has been implemented by Matthias Zschunke (Transinsight GmbH).

GoPubMed

Three years of design and implementation and of various modules, as i.e. the Author Network Visualization, the Top Terms algorithm, Grammar-based query parser, as well as other underlying software components together with Heiko Dietze, Andreas Doms, Loic A. Royer (TU Dresden), and Transinsight GmbH.

⁶ <http://www.hicknhack-software.com>

LMOPubMed

Design, implementation, and evaluation of the LMOPubMed semantic search engine. This included the work on the ontology together with Dimitra Alexopoulou (TU Dresden) and the collaboration with Transinsight GmbH (Dresden) and Unilever Research (Colworth, UK).

MousePubMed

Design, implementation, and evaluation of the MousePubMed search engine and in particular matching algorithms for developmental stages in early mouse development with Dr. Jörg Hakenberg and the disambiguation method for highly ambiguous concepts in anatomy.

Prototypes

Development of various prototypes to explore different document sources and ontologies.

- **GoGoogle** – browsing Google search results with Gene Ontology
- **FSTAPubMed** – browsing PubMed abstracts with the taxonomy of the Food Science and Technology Abstracts Database.
- **ZebetPubMed** and **ZebetZebet** – browsing PubMed and ZEBET abstracts with the initial ZEBET ontology (predecessor of Go3R)
- **GoCell** – indexing 10 years of full text articles from the Elsevier Journal Cell with the Gene Ontology

All prototypes have been build on the basis of early versions of GoPubMed.

Conclusion and Future Work

This chapter summarises the solutions to the open research questions listed in Chapter 1 (Introduction) and the scientific contributions to the semi-automated development of biomedical ontologies, addressing the needs of biocuration as well as the application of ontologies in the first semantic search engine for alternative methods to animal testing.

9.1 Semi-automated ontology generation

Research Question 1

To what extent can ontology construction be automated?

The goal of this work is to design and to implement ontology generation methods for the generation of terms, definitions, and taxonomic relationships in the life sciences. The methods should be fast and scalable to be suitable for integration in interactive applications. A thorough evaluation is required to build up user acceptance and to allow estimations on the overall quality of the methods.

On the basis of a comprehensive survey on the state-of-the-art of automatic term recognition, abbreviation detection, definition extraction, definitional question answering, and taxonomy generation new ontology generation methods have been developed. Each new method was systematically evaluated at a large scale using manually validated benchmarks. The results obtained in experiments in the Chapters 3, 4, and 5 show that text-mining supports ontology engineers with highly relevant terms, definitions, and parent-child relations.

Term generation The developed method generates ranked lists of terms by identifying statistically significant noun-phrases in text and it specifically deals with the relevance ranking of single word terms by using reference corpora and term normalisation. Despite the additional difficulty retrieving and ranking single word terms, the method performs equally well or better than the available systems TerMine, Text2Onto, and OntoLearn. The method is capable of retrieving 75% relevant terms in the top 50 suggestions in 1-2 seconds which is sufficient for integration in interactive applications. Among the generated terms over 80% can be mapped

to terms in other ontologies which shows that the notion of statistically significant noun phrases is a good approximation for manually defined term labels. This high amount of terms that already exist is a strong motivation for the re-use of existing resources before creating new terms.

Stability of the term ranking The stability of the relevance ranking depends on the source of the global corpus statistics as well as on the used scoring method. An alternative Part-Of-Speech tagger for noun phrase extraction only marginally affected the position at which terms were retrieved.

The term weighting in contrast to a large global reference corpus led to the promotion of relevant terms. Against the hypothesis the general occurrence counts obtained from a large set of web sites performed equally well or even better than domain specific occurrence counts obtained from PubMed. Hence, the method is domain independent and directly applicable to other domains without the need of a domain specific reference corpus. It has been confirmed in the experiments that the scoring weight tf-idf (term frequency-inverse document frequency) is an efficient approximation for the conditional probability of occurrence given the frequency counts from a global reference corpus, but tf-idf promotes many terms which have not been ranked particularly high based on true conditional probability measurements. Therefore the true probabilities should be used provided they are efficiently calculated as proposed with the approximation described in Section 6.3.3.

Definition extraction A definition extraction method has been developed and evaluated. The method is capable of extracting and ranking relevant definitional sentences from web search results and web sites. Each generation step was systematically evaluated using manually validated benchmarks. Definitions have been extracted for 500 randomly selected GO and MeSH terms. The top 10 ranked definitions per term, in total 10,000 generated definitions, were manually evaluated to find the best ranked correct or good alternative definitions. For 32% of terms the first extracted definition was correct compared to the original term definition and for 47% of terms it was at least good, but different to the terms original definition. For up to 78% of terms good definitions could be retrieved in the top 10 ranked definitions. No other validated system exists that achieves comparable results.

Taxonomy generation A method to generate taxonomic relations on the basis of generated definitions has been developed and evaluated for 1,000 randomly selected ontology terms from GO and MeSH. For 38% (GO) and 54% (MeSH) of the terms, correct relations to ancestors could be predicted. For the majority of terms the first retrieved definitions contained a correct relation. Thus, definitions are a high quality source for taxonomic relations and for a significant number of current ontology terms the relations to parents can be found in definitions extracted from the Web.

Two additional experiments have been performed to test the suitability of pattern-based and statistical methods for finding parent-child relations using the super string property, namely the inclusion of the parent term in the child term and the co-occurrence of terms in very large corpora. Based on a data set of approximately 200,000,000 term occurrences in PubMed abstracts it has been experimentally analysed how well the taxonomic relations of the MeSH can be reconstructed.

Overall results ranged around an F-measure of 0.15 but parts of MeSH could be reconstructed with an F-measure greater than 0.6.

The methods have been encapsulated in web services which enabled the integration in various applications.

9.2 Automated ontology generation support for biocuration

Research Question 2

How can ontology generation methods be integrated into ontology editors?

The goal is to find solutions to integrate ontology generation methods into existing ontology editors used for the development of bio-ontologies. This includes the possibility to generate terms, textual definitions, and taxonomic relations, as well as the re-use of existing ontologies.

The *Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG)* developed for this thesis (Chapter 7) is a system which supports the creation and extension of ontologies by semi-automatically generating terms, definitions, and parent-child relations from text in PubMed, the Web, and PDF repositories. The system is seamlessly integrated into OBO-Edit and Protégé, two widely used ontology editors in the life sciences.

In the Gene Ontology annotation process described by Hill et al. (2008), *DOG4DAG* can help to identify appropriate ontology annotation terms and literature references as evidence to include in the annotation record. In cases where novel terms need to be created *DOG4DAG* helps to define and place the new term in the ontology.

Definitions of terms in ontologies are important, but cumbersome to define. In the over 90 OBO ontologies there are 99,418 terms without a definition. Thus, there is a huge potential to save work time when defining terms with *DOG4DAG*.

By combining the prediction of high quality terms, definitions and parent-child relations with the ontology editors, two thoroughly validated tools have been contributed for all ontology engineers in the life sciences and beyond.

9.3 Semantic search for alternative methods to animal testing

Research Question 3

How to employ ontology generation methods and ontology-based search to determine the availability or unavailability of alternative methods to animal testing.

Currently, there is no ontology for alternative methods to animal testing. How can such an ontology be created using editing tools? How applicable are automated methods for ontology generation as discussed in research question 1 and made available as answer to question 2? Finally, how can such an ontology be used to improve the search for information relevant to the 3Rs principle (Russell and Burch, 1959) and what are the limits of such an approach?

Chapter 8 described the design and implementation of Go3R, the first and currently only semantic tool with a specific focus on alternative methods to animal testing.

Ontology development: An ontology for alternative methods to animal testing has been developed which comprises in total 17,151 terms and 70,840 synonyms including Diseases & Symptoms, Body Systems & Structures, and Statistics from MeSH. 1,779 terms and 1,419 synonyms have been newly defined to capture the domain relevant vocabulary for alternative methods. To facilitate the efficient creation of the taxonomic backbone of ontologies a user friendly editor has been newly designed (Section 7.4).

Ontology generation: Experiments showed that the developed methods and tools for semi-automated ontology generation can contribute to the future extension and maintenance of the ontology by suggesting terms and enriching their lexical base for a better recognition in text. The definition generation method can find definitions for half of the 152 alternative methods for toxicity testing which exist in Go3R.

Ontology-based search: The web-based search engine provides the search results automatically linked to terms of the Go3R ontology which serves as “intelligent table of contents”. Thereby, Go3R actively supports the user in finding information on alternative methods by automatically classifying documents in accordance with the 3Rs principle (Russell and Burch, 1959) as 3Rs relevant in general and as relevant to the replacement, reduction and refinement of animal experiments in particular. For this, Go3R employs machine learning techniques based on user feedback in form of relevance curations. A first evaluation showed high (> 90%) classification accuracy for determining the relevance of a document for the domain. An integrated process of user feedback, disambiguation, and immediate application was set up to ensure long term quality of Go3R search results.

Go3R aims to enable all those involved in the planning, authorisation, and performance of animal experiments to determine the existence of non-animal methodologies in a fast, comprehensive, and transparent manner. Future assessments during the ongoing BMBF funded project “Entwicklung und Etablierung einer Semantischen Suchmaschine für Alternativmethoden zu Tierversuchen (FK:614 40003

0315489A)” which resulted from this work will show to what extent Go3R can contribute to a reduction the number of animals used in toxicity testing in industry.

9.4 Future Work

Ontology generation and tools

The ontologies created for this thesis are defined in close relation to text and are intended for document annotation and document indexing. The majority of ontologies in the life sciences belong to this category. Nevertheless formal ontologies with richer semantics exist and are used. Examples are SNOMED in Medicine, the Foundational Model of Anatomy (Rosse and Mejino, 2003), or CHEBI in Chemistry (de Matos et al., 2010). It remains open if and to what extent nontrivial ontologies with several relationship types and logical definitions can be automatically constructed.

The great amount of generated terms that were already defined in the UMLS (Section 3.4.1) motivates the consistent re-use of terms and relationships as primary source for semi-automated ontology in the life sciences. Future work will have to discover, how the correct portion of “ontology” can be extracted from the available resource, how the divers modelling primitives used in different ontologies can be overcome, and how it can be ensured that new developments in the original sources find their way into the automatically derived offspring.

Go3R research project

In the ongoing BMBF funded research project Go3R will be extended to find toxicological and hazard data needed for substance registration under REACH and for general safety assessment. All information found on the hazard potential of a substance or suitable measurement techniques, can potentially reduce the required number of animal experiments.

Ontology-based search

User feedback over the last five years showed that navigating data sets using the taxonomic backbone of ontologies is generally powerful in itself, but still too complicated for the majority of intended users. After showing that meaningful comprehensive results can be achieved by incorporating ontologies as background knowledge in search engines, further efforts need to be made to hide this explicit structure and only reveal it where appropriate. This requires first, to create the semantic awareness to be able to understand what a user has asked for, second to rank and filter possible answers, and third, to explain to the user how and why this specific answer has been selected.

References

- Adar, E. (2004). SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics (Oxford, England)*, 20(4):527–533.
- Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., and Wächter, T. (2009). Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics*, 10:28.
- Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C., and Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*, 9(Suppl 9):S2.
- Altun, Z. F. and Hall, D. H., editors (2002-2006). *WormAtlas*.
- Ando, R. K. (2007). BioCreative II Gene Mention Tagging System at IBM Watson. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 101–103.
- Ao, H. and Takagi, T. (2005). ALICE: An Algorithm to Extract Abbreviations from MEDLINE. *J Am Med Inform Assoc*, 12(5):576–586.
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2004). Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue).
- Aranguren, M., Antezana, E., Kuiper, M., and Stevens, R. (2008). Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. *BMC Bioinformatics*, 9(Suppl 5):S1.
- Arpe, A. (1995). Term extraction from unrestricted text. In *10th Nordic Conference of Computational Linguistics (NoDaLiDa)*.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29.
- Azuma, N., Tadokoro, K., Asaka, A., Yamada, M., Yamaguchi, Y., Handa, H., Matsushima, S., Watanabe, T., Kida, Y., Ogura, T., Torii, M., Shimamura, K., and Nakafuku, M. (2005). Transdifferentiation of the retinal pigment epithelia to the neural retina by transfer of the pax6 transcriptional factor. *Hum Mol Genet*, 14(8):1059–68.
- Baclawski, K. and Niu, T. (2005). *Ontologies for Bioinformatics (Computational Molecular Biology)*. The MIT Press.
- Baldock, R. A., Bard, J. B. L., Burger, A., Burton, N., Christiansen, J., Feng, G., Hill, B., Houghton, D., Kaufman, M., Rao, J., Sharpe, J., Ross, A., Stevenson, P., Venkataraman, S., Waterhouse, A., Yang, Y., and Davidson, D. R. (2003). EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics*, 1(4):309–25.
- Bard, J., Rhee, S. Y., and Ashburner, M. (2005). An ontology for cell types. *Genome Biol*, 6(2).
- Bard, J. L., Kaufman, M. H., Dubreuil, C., Brune, R. M., Burger, A., Baldock, R. A., and Davidson, D. R. (1998). An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech Dev*, 74(1-2):111–20.
- Bechhofer, S., Horrocks, I., Goble, C. A., and Stevens, R. (2001). Oiled: A reason-able ontology editor for the semantic web. In *KI/Ā-GAL*, pages 396–408.
- Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N.,

II References

- Weems, D., and Rhee, S. Y. (2004). Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiol*, 135(2):745–755.
- Berger, A. L., Della Pietra, S. D., and Della Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Comp. Linguistics*, 22(1):39–71.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.
- Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., and Eppig, J. T. a. (2003). Mgd: the mouse genome database. *Nucleic Acids Res*, 31(1):193–195.
- BMELV (2008). German Federal Ministry for Nutrition, Agriculture and Consumer Protection: Numbers of Animals Used in Scientific Procedures in 2008. Tierversuchszahlen 2008. <http://www.bmelv.de/cae/servlet/contentblob/765788/publicationFile/43424/2008-TierversuchszahlenGesamt.pdf>. [Online; accessed March-2010].
- Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue).
- Bodenreider, O., Burgun, A., and Mitchell, J. A. (2003). Evaluation of wordnet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. *Stud Health Technol Inform*, 95:379–384.
- Bodenreider, O. and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256.
- Bontas, E. P., Tempich, C., and Sure, Y. (2006). Ontocom: A cost estimation model for ontology engineering. In Cruz, I. et al., editors, *Proc of the 5th Int Semantic Web Conf (ISWC 2006)*, volume 4273 of *Lecture Notes in Computer Science (LNCS)*, pages 625–639. Springer-Verlag Berlin Heidelberg.
- Bourigault, D. (1994). *LEXTER, un Logiciel d'Extraction de TERminologie. Application a l'acquisition des connaissances a partir de textes*. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.
- Bourne, P. E. and McEntyre, J. (2006). Biocurators: Contributors to the world of science. *PLoS Comput Biol*, 2(10):e142.
- Brants, T. (2000). Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231, Morristown, NJ, USA. Association for Computational Linguistics.
- Brants, T. and Franz, A. (2006). Web 1t 5-gram version 1. Linguistic Data Consortium, Philadelphia.
- Brewster, C., Jupp, S., Luciano, J., Shotton, D., Stevens, R. D., and Zhang, Z. (2009). Issues in learning an ontology from text. *BMC Bioinformatics*, 10 Suppl 5:S1.
- Buitelaar, P., Declerck, T., Sacaleanu, B., Vintar, S., Raileanu, D., and Crispi, C. (2003). A multi-layered, xml-based approach to the integration of linguistic and semantic annotations. In *Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web*.
- Buitelaar, P., Olejnik, D., and Sintek, M. (2004). A Protégé plug-in for ontology extraction from text based on linguistic analysis. In *The Semantic Web: Research and Applications*, volume 3053 of *LNCS*, pages 31–44. Springer, Berlin / Heidelberg.
- Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Morristown, NJ, USA. Association for Computational Linguistics.
- Carpenter, B. (2009). Alias-i. <http://alias-i.com/lingpipe> (accessed July 27, 2009).
- Castellví, M. T., Bagot, R. E., and Palatresi, J. V. (2001). Automatic term detection: A review of current systems. In Bourigault, D., Jacquemin, C., and L'Homme, M.-C., editors, *Recent Advances in Comp Terminology*, pages 53–88. John Benjamins, Amsterdam/Philadelphia.
- Castori, M., Barboni, L., Duncan, P. J., Paradisi, M., Laino, L., De Bernardo, C., Robinson, D. O., and Grammatico, P. (2009). Darier disease, multiple bone cysts, and aniridia due to double de novo heterozygous mutations in atp2a2 and pax6. *Am J Med Genet A*, 149A(8):1768–72.
- Chang, J. T. and Schütze, H. (2006). *Text Mining for Biology and Biomedicine.*, chapter Abbreviations in biomedical text., pages 99–119. Artech House Inc, London.
- Chang, J. T., Schütze, H., and Altman, R. B. (2002). Creating an online dictionary of abbreviations from medline. *J Am Med Inform Assoc*, 9(6):612–620.
- Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). Sgd: Saccharomyces genome database. *Nucleic Acids Res*, 26(1):73–79.
- Chiang, J. H. and Yu, H. C. (2004). Extracting functional annotations of proteins based on hybrid text mining approaches. In *Proc. of BioCreative*.

- Chomsky, N. (1957). *Syntactic Structures*. Mouton and Co, The Hague.
- Cimiano, P. (2006a). *Ontology Learning and Population from Text*. PhD thesis, PhD thesis at the Universität Karlsruhe (TH), Fakultät für Wirtschaftswissenschaften.
- Cimiano, P. (2006b). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer.
- Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res.*, 24(1):305–339.
- Cimiano, P. and Völker, J. (2005). Text2Onto - a framework for ontology learning and data-driven change discovery. In Montoyo, A., Munoz, R., and Metais, E., editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *LNCS*, pages 227–238. Springer, Alicante, Spain.
- Collier, N., Nobata, C., and Tsujii, J.-i. (2000). Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th conference on Computational linguistics*, pages 201–207, Morristown, NJ, USA. Association for Computational Linguistics.
- Commission of the European Communities (1986). Council directive 86/609/eec of 24 november 1986 on the approximation of laws, regulations and administrative provisions of the member states regarding the protection of animals used for experimental and other scientific purposes. Official Journal L 358, 1. 28 December 1986. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31986L0609:EN:HTML>. [Online; accessed January-2008].
- Commission of the European Communities (1991). Council directive 91/414/eec of 15 july 1991 concerning the placing of plant protection products on the market. Official Journal L 230, 19 August 1991, p. 1-32. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31991L0414:EN:HTML>.
- Commission of the European Communities (2006). Regulation (ec) no 1907/2006 of the european parliament and of the council of 18 december 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (reach), establishing a european chemicals agency, amending directive 1999/45/ec and repealing council regulation (eec) no 793/93 and commission regulation (ec) no 1488/94 as well as council directive 76/769/eec and commission directives 91/155/eec, 93/67/eec, 93/105/ec and 2000/21/ec. Official Journal L 396, 30 December 2006, p. 1-849. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:396:0001:0849:EN:PDF>.
- Commission of the European Communities (2008). Proposal for a directive of the european parliament and of the council on the protection of animals used for scientific purposes. COM/2008/0543 final - COD 2008/0211, 5 November 2008. 88 pp. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0543:FIN:EN:PDF>.
- Couto, F., Silva, M., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., and Rebholz-Schuhmann, D. (2006). GOAnnotator: linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1(1):19.
- Couto, F. M., Silva, M. J., and Coutinho, P. M. (2005). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6 Suppl 1.
- Cui, H., Kan, M.-Y., and Chua, T.-S. (2004). Unsupervised learning of soft patterns for generating definitions from online news. In *Proc of WWW04*, pages 90–99, New York, NY, USA. ACM.
- Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the ACL*.
- Day-Richter, J., Harris, M. A., Haendel, M., and Lewis, S. (2007). Obo-edit—an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200.
- de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fritzsche, F., Walther, T. C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–4.
- de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical entities of biological interest: an update. *Nucleic Acids Res*, 38(Database issue):D249–D254.
- Degórski, L., Marcińczuk, M., and Przepiórkowski, A. (2008). Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association.

- Dietze, H., Alexopoulou, D., Alvers, M. R., Barrio-Alvers, B., Doms, A., Hakenberg, J., Mönnich, J., Plake, C., Reischuk, A., Royer, L., Wächter, T., Zschunke, M., and Schroeder, M. (2008a). *GoPubMed: Exploring Pubmed with Ontological Background Knowledge*, chapter Bioinformatics for Systems Biology. Humana Press.
- Dietze, H., Alexopoulou, D., Alvers, M. R., Barrio-Alvers, B., Doms, A., Hakenberg, J., Mönnich, J., Plake, C., Reischuk, A., Royer, L., Wächter, T., Zschunke, M., and Schroeder, M. (2008b). *GoPubMed: Exploring Pubmed with Ontological Background Knowledge*. In Ashburner, M., Leser, U., and Rebholz-Schuhmann, D., editors, *Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives*, number 08131 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.
- Dietze, H. and Schroeder, M. (2008). GoWeb: A semantic search engine for the life science web. In Albert Burger, Adrian Paschke, P. R. and Splendiani, A., editors, *In Proceedings of the Intl. Workshop on Semantic Web Applications and Tools for the Life Sciences SWAT4LS*.
- Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., and Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, 39(10):1202–7.
- Doms, A. (2004). Using sequence alignment algorithms to extract gene ontology terms in biomedical literature abstracts. Diplomathesis, TU Dresden.
- Doms, A. (2009). *GoPubMed: Ontology-based literature search for the life sciences*. PhD thesis, Technische Universität Dresden.
- Doms, A., Jakoniene, V., Lambrix, P., Schroeder, M., and Wächter, T. (2006). Ontologies and Text Mining as a Basis for a Semantic Web for the Life Sciences. In Barahona, P., Bry, F., Franconi, E., Henze, N., and Sattler, U., editors, *Reasoning Web*, volume 4126 of *Lecture Notes in Computer Science*, pages 164–183. Springer.
- Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, 33(Web Server issue).
- Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J., and Blake, J. A. (2009). Integrating text mining into the mgi biocuration workflow. *Database : the journal of biological databases and curation*, 2009(0):bap019+.
- Echihabi, A., Hermjakob, U., Hovy, E. H., Marcu, D., Melz, E., and Ravichandran, D. (2003). Multiple-engine question answering in TextMap. In *Proceedings of the 12th Text Retrieval Conference (TREC-2003)*, pages 772–781, Gaithersburg, Maryland.
- Eilbeck, K., Lewis, S. E., Mungall, C., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biol*, 6(5):R44.
- Evans, D. A. and Zhai, C. (1996). Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 17–24, Morristown, NJ, USA. Association for Computational Linguistics.
- Faure, D. and N’edellec, C. (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*, pages 5–12, Granada, Spain.
- Faure, D. and N’edellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In *EKAU ’99: Proc of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*, pages 329–334, London, UK. Springer-Verlag.
- Faure, D. and Poibeau, T. (2000). First experiments of using semantic knowledge learned by asium for information extraction task using intex. In S. Staab, A. Maedche, C. Nédellec, P. Wiemer-Hastings(eds.), *Proc of the Workshop on Ontology Learning, 14th European Conf on Artificial Intelligence ECAI00, Berlin, Germany*.
- Fauser, B. (2007). Parliamentary petition of mp beate fauser and others and statement thereupon by the minister for the environment regarding the reach regulation. parliament of the federal state of baden-wuerttemberg. - antrag der abg. beate fauser u. a. fdp/dvp und stellungnahme des umwelt-ministeriums zur reach chemikalienverordnung, landtag von baden-württemberg. Drucksache 14 / 1166 14. Wahlperiode, 19. 04. 2007.
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Flybase Consortium (1999). The flybase database of the drosophila genome projects and community literature. the flybase consortium. *Nucleic Acids Res*, 27(1):85–88.

- Frantzi, K. and Ananiadou (1995). Statistical measures for terminological extraction. Technical report, Department of Computing of Manchester Metropolitan University.
- Frantzi, K. and Ananiadou, S. (1997). Automatic term recognition using contextual clues. In *Proc of Mulsaic 97, IJCAI, Japan*. 8.
- Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.
- Frantzi, K. T., Ananiadou, S., and ichi Tsujii, J. (1998). The C-value/NC-value method of automatic recognition for multi-word terms. In *ECDL '98: Proc of the Second European Conf on Research and Advanced Technology for Digital Libraries*, pages 585–604, London, UK. Springer-Verlag.
- Fukuda, K. (1998). Toward information extraction: identifying protein names from biological papers. In *Proc. of the Pacific Symposium on Biocomputing*, pages 707–718.
- Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA. Association for Computational Linguistics.
- Ganter, B., Stumme, G., and Wille, R., editors (2005). *Formal Concept Analysis, Foundations and Applications*, volume 3626. Springer.
- Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005). Resolving abbreviations to their senses in MEDLINE. *Bioinformatics*, 21(18):3658–64.
- Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 6(2):199–221.
- Grumblin, G. and Strelets, V. (2006). FlyBase: anatomical data, images and queries. *Nucleic Acids Res*, 34(Database issue):D484–8.
- Grune, B., Fallon, M., Howard, C., Hudson, V., Kulpa-Eddy, J. A., Larson, J., Leary, S., Roi, A., van der Valk, J., Wood, M., Dörendahl, A., Köhler-Hahn, D., Box, R., Spielmann, H., and approaches for information on alternative methods to animal experiments, R. (2004). Report and recommendations of the international workshop "retrieval approaches for information on alternative methods to animal experiments". *ALTEX*, 21(3):115–27.
- Haase, P. and Stojanovic, L. (2005). Consistent evolution of owl ontologies. In *Proceedings of the Second European Semantic Web Conference, Heraklion, Greece, 2005*, volume 3532 of LNCS, pages 182–197. Springer.
- Hagiwara, M., Ogawa, Y., and Toyama, K. (2006). Selection of effective contextual information for automatic synonym acquisition. In *Proc of COLING06*, pages 353–360, Sydney, Australia. ACL.
- Hahn, U. and Schnattinger, K. (1998). Towards text knowledge engineering. In *AAAI/IAAI*, pages 524–531.
- Hakenberg, J. (2007). What's in a gene name?: automated refinement of gene name dictionaries. In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, pages 153–160, Morristown, NJ, USA. Association for Computational Linguistics.
- Hakenberg, J., Plake, C., Royer, L., Strobelt, H., Leser, U., and Schroeder, M. (2008). Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*, 9 Suppl 2:S14–S14.
- Hakkinen, P. J. B. and Green, D. K. (2002). Alternatives to animal testing: information resources via the internet and world wide web. *Toxicology*, 173(1-2):3–11.
- Han, K.-S., Song, Y.-I., Kim, S.-B., and Rim, H.-C. (2006). A definitional question answering system based on phrase extraction using syntactic patterns. *IEICE - Trans. Inf. Syst.*, E89-D(4):1601–1605.
- Harris, Z. (1968). *Mathematical Structures of Language*. Wiley.
- Hartung, T. and Rovida, C. (2009). Chemical regulators have overreached. *Nature*, 460:1080–1081.
- Hatala, M., Gasevic, D., Siadat, M., Jovanovic, J., and Torniai, C. (2009). Utility of ontology extraction tools in the hands of educators. In *ICSC '09: Proceedings of the 2009 IEEE International Conference on Semantic Computing*, pages 408–413, Washington, DC, USA. IEEE Computer Society.
- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.
- Henschel, A., Wong, W. L., Wächter, T., and Madnick, S. (2009). Comparison of generality based algorithm variants for automatic taxonomy generation. In *6th International Conference on Innovations in Information Technology*, AlAin, United Arab Emirates.
- Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University.

- Hill, D., Smith, B., McAndrews-Hill, M., and Blake, J. (2008). Gene ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9(Suppl 5):S2.
- Hirschman, L., Morgan, A. A., and Yeh, A. S. (2002). Rutabaga by any other name: extracting biological names. *J. of Biomedical Informatics*, 35(4):247–259.
- Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1.
- Hisamitsu, T. and Niwa, Y. (2001). *Recent Advances in Computational Terminology*, chapter Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures: A comparative evaluation of bigram statistics., pages 209–224. Natural Language Processing. John Benjamins, Amsterdam.
- Höfer, T., Gerner, I., Gundert-Remy, U., Liebsch, M., Schulte, A., Spielmann, H., Vogel, R., and Wettig, K. (2004). Animal testing and alternative approaches for the human health risk assessment under the proposed new european chemicals regulation. *Arch Toxicol*, 78(10):549–64.
- Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nat Genet*, 36(7).
- Hoffmann, R. and Valencia, A. (2005). Implementing the ihop concept for navigation of biomedical literature. *Bioinformatics*, 21(2):252–258.
- Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47–50.
- Hsieh, Y. W. and Yang, X. J. (2009). Dynamic pax6 expression during the neurogenic cell cycle influences proliferation and cell fate choices of retinal progenitors. *Neural Dev*, 4(1):32.
- Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B. W., Beavis, W., Meinke, D. W., Town, C. D., Somerville, C., and Rhee, S. Y. (2001). The arabidopsis information resource (tair): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*, 29(1):102–105.
- Huang, H.-S., Lin, Y.-S., Lin, K.-T., Kuo, C.-J., Chang, Y.-M., Yang, B.-H., Chung, I.-F., and Hsu, C.-N. (2007). High-recall gene mention recognition by unification of multiple backward parsing models. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 109–111.
- Humphreys, B., Lindberg, D., Schoolman, H., and Barnett, G. (1998). The unified medical language system: an informatics research collaboration. *J Am Med Inform Assoc*, 5(1):1–11.
- Jacquín, C. and Liscouet, M. (1996). Terminology extraction from texts corpora: application to document keeping via internet. In *TKE '96: Terminology and Knowledge Engineering*, pages 74–83. Indeks Verlag Berlin.
- Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., and Zapata, F. (2005). Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comp Funct Genomics*, 6(7-8):388–97.
- Jarmasz, M. and Szpakowicz, S. (2003). Roget’s thesaurus and semantic similarity. In *Proc of Conf on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 212–219.
- Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2):119–129.
- Justeson, J. and Katz, S. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.
- Klavans, J. L. and Muresan, S. (2000). DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In *Proc AMIA Symp.*, page 1049.
- Klavans, J. L. and Muresan, S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proc AMIA Symp.*, pages 324–28.
- Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *J. of Biomed. Informatics*, 37(6):512–26.
- Kuo, C.-J., Chang, Y.-M., Huang, H.-S., Lin, K.-T., Yang, B.-H., Lin, Y.-S., Hsu, C.-N., and Chung, I.-F. (2007). Rich feature set, unification of bidirectional parsing and dictionary filtering for high f-score gene mention tagging. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 105–107.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

- Lee, J. B., Kim, J. J., and Park, J. C. (2006). Automatic extension of Gene Ontology with flexible identification of candidate terms. *Bioinformatics*, 22(6):665–70.
- Liu, B., Chin, C. W., and Ng, H. T. (2003). Mining topic-specific concepts and definitions on the web. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 251–260, New York, NY, USA. ACM.
- Liu, H., Aronson, A. R., and Friedman, C. (2002). A study of abbreviations in MEDLINE abstracts. *Proc AMIA Symp*, pages 464–468.
- Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems and Their Applications*, 16(2):72–79.
- Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 251–263, London, UK. Springer-Verlag.
- Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 33(Database issue):D54–D58.
- Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.
- Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330.
- Mccrae, J. and Collier, N. (2008). Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, 9:159+.
- Morgan, A. and Hirschman, L. (2007). Overview of BioCreative II Gene Normalization. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 17–27.
- Mosler, K. and Schmid, F. (2006). *Wahrscheinlichkeitsrechnung und schließende Statistik*, chapter Stichproben und Stichprobenfunktionen, pages 173–193. Springer-Lehrbuch. Springer.
- Müller, H. M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309.
- Mungall, C., Gkoutos, G., Smith, C., Haendel, M., Lewis, S., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+.
- Mungall, C. J. (2004). Obol: integrating language and meaning in bio-ontologies: Conference Papers. *Comp Funct Genomics*, 5(6-7):509–520.
- Nadeau, D. and Turney, P. D. (2005). A supervised learning approach to acronym identification. In *8th Canadian Conference on Artificial Intelligence (AI'2005) (LNAI 3501)*, pages 319–329.
- Natale, D. A., Arighi, C. N., Barker, W., Blake, J., Chang, T.-C., Hu, Z., Liu, H., Smith, B., and Wu, C. H. (2006). Framework for a protein ontology. *BMC Bioinformatics*, 8(Suppl 9):29–36.
- Navigli, R. and Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.*, 30(2):151–179.
- Navigli, R. and Velardi, P. (2006). Enriching a formal ontology with a thesaurus: an application in the cultural heritage domain. In *Proc of the ACL-2nd Workshop on Ontology Learning and Population*, pages 1–9, Sydney, Australia. ACL.
- Navigli, R., Velardi, P., Cucchiarelli, A., and Neri, F. (2004). Quantitative and qualitative evaluation of the ontolearn ontology learning system. In *Proc of COLING04*, page 1043, Morristown, USA. ACL.
- Nenadić, G., Ananiadou, S., and McNaught, J. (2004a). Enhancing automatic term recognition through recognition of variation. In *Proc of COLING04*, page 604, Morristown, USA. ACL.
- Nenadić, G., Spasic, I., and Ananiadou, S. (2004b). Mining biomedical abstracts: What is in a term? In *Proc of Int Joint Conf on NLP*, pages 247–254, Sanya, China.
- Nicholson, A., Sandler, J., and Seidle, T. (2004). An evaluation of the us high production volume (hpv) chemical-testing programme: A study in (ir)relevance, redundancy and retro thinking. *ATLA. Alternatives to laboratory animals*, 32:335–341. SUP1A.
- Ogren, P. V., Cohen, K. B., Acquah-Mensah, G. K., Eberlein, J., and Hunter, L. (2004). The compositional structure of Gene Ontology terms. *Pacific Symposium on Biocomputing*, pages 214–225.
- Ohta, T., Tateisi, Y., and Kim, J.-D. (2002). The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proc of the 2nd Int Conf on Human Language Technology Research*, pages 82–86, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Okazaki, N. and Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–95.

- Okazaki, N., Ananiadou, S., and Tsujii, J. (2008). A discriminative alignment model for abbreviation recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 657–664, Manchester, UK. Coling 2008 Organizing Committee.
- Pakhomov, S. (2001). Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.
- Patel-Schneider, P. F. and Horrocks, I. (2004). OWL web ontology language semantics and abstract syntax.
- Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proc of the ACL*, pages 183–190, Morristown, USA. ACL.
- Perez-Iratxeta, C., Pérez, A. J., Bork, P., and Andrade, M. A. (2003). Update on xplormed: A web server for exploring scientific literature. *Nucleic Acids Res*, 31(13):3866–3868.
- Pivk, A. (2006). Thesis: automatic ontology generation from web tabular structures. *AI Commun.*, 19(1):83–85.
- Plake, C., Royer, L., Winnenburger, R., Hakenberg, J., and Schroeder, M. (2009). Gogene: gene annotation in the fast lane. *Nucleic Acids Research*, 37(Web-Server-Issue):300–304.
- Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., and Leser, U. (2006). Alibaba: Pubmed as a graph. *Bioinformatics*, 22(19):2444–2445.
- Porter, M. F. (1997). An algorithm for suffix stripping. pages 313–316.
- Przepiórkowski, A., Degórski, L., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kubon, V., and Wójtowicz, B. (2007). Towards the automatic extraction of definitions in slavic. In *Proc of the ACL-Workshop on Balto-Slavonic Natural Language Processing*, Prag.
- Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M., and Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from medline databases. *Stud Health Technol Inform*, 84(Pt 1):371–375.
- Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In Yarovsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics.
- Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 41–47, Morristown, NJ, USA. Association for Computational Linguistics.
- Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., and Stoehr, P. (2007). Ebimed–text crunching to gather facts for proteins from medline. *Bioinformatics*, 23(2).
- Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–5.
- Rosse, C. and Mejino, J. L. V. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36(6):478–500.
- Russell, W. and Burch, R. (1959). *he Principles of Humane Experimental Technique*. Methuen, London.
- Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M. S., Ogbuji, C., Rees, J., Stephens, S., Wong, G., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., and Cheung, K.-H. (2007). Advancing translational research with the semantic web. *BMC Bioinformatics*, 8(Suppl 3):S2.
- Ryu, P.-M. and Choi, K.-S. (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, chapter An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning. Frontiers in Artificial Intelligence and Applications, ios press edition.
- Ryu, P.-M. and Choi, K.-S. (2006). Taxonomy learning using term specificity and similarity. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 41–48, Sydney, Australia. Association for Computational Linguistics.
- Saggion, H. and Gaizauskas, R. J. (2004). Mining on-line sources for definition knowledge. In Barr, V. and Markov, Z., editors, *FLAIRS Conf*. AAAI Press.
- Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA. ACM.
- Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., and Albrecht, M. (2007). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7):859–65.

- Schober, D., Smith, B., Lewis, S., Kusnierczyk, W., Lomax, J., Mungall, C., Taylor, C., Rocca-Serra, P., and Sansone, S.-A. (2009). Survey-based naming conventions for use in obo foundry ontology development. *BMC Bioinformatics*, 10(1):125.
- Schuemie, M. J., Kors, J. A., and Mons, B. (2005). Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol*, 12(5):554–565.
- Schwartz, A. S. and Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 451–462.
- Sclano, F. and Velardi, P. (2007). Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proc of the 3rd Int Conf on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal (Madeira Island), Portugal.
- Shimizu, N., Hagiwara, M., Ogawa, Y., Toyama, K., and Nakagawa, H. (2008). Metric learning for synonym acquisition. In *Proc of COLING08*, pages 793–800, Manchester, UK. ACL.
- Shimohata, M. and Sumita, E. (2005). Acquiring synonyms from monolingual comparable texts. In Dale, R., Wong, K.-F., Su, J., and Kwong, O. Y., editors, *IJCNLP*, volume 3651 of *LNCS*, pages 233–244. Springer.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.
- Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., and Rosse, C. (2005a). Relations in biomedical ontologies. *Genome Biology*, 6(5):R46.
- Smith, C. L., Goldsmith, C. A., and Eppig, J. T. (2005b). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1).
- Smith, L., Rindflesch, T., and Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.
- Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.
- Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304, Cambridge, MA. MIT Press.
- Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogeneous evidence. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808, Morristown, NJ, USA. Association for Computational Linguistics.
- Snow, R., Prakash, S., Jurafsky, D., and Ng, A. Y. (2007). Learning to Merge Word Senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1005–1014.
- Soldatova, L. N. and King, R. D. (2005). Are the current ontologies in biology good ontologies? *Nature Biotechnology*, 23(9):1095–1098.
- Spasić, I., Schober, D., Sansone, S. A., Rebholz-Schuhmann, D., Kell, D. B., and Paton, N. W. (2008). Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*, 9(Suppl 5).
- Stojanovic, L. and Motik, B. (2002). Ontology evolution within ontology editors. In *In Proc of the OntoWeb-SIG3 Workshop at the 13th Int Conf 22 on Knowledge Engineering and Knowledge Management (EKAW)*, pages 53–62.
- Storrer, A. and Wellinghoff, S. (2006). Automated detection and annotation of term definitions in german text corpora. In *Proc of fifth international conference on Language Resources and Evaluation (LREC)*, Genua, Italy.
- Taghva, K. and Gilbreth, J. (1999). Recognizing acronyms and their definitions. *ISRI (Information Science Research Institute) UNLV*, 1:191–198.
- Tanabe, L. and Wilbur, W. J. (2002). Tagging gene and protein names in full text articles. In *Proc of the ACL-workshop on Natural language processing in the biomedical domain*, pages 9–13, Morristown, USA. ACL.
- Taylor, D. P. (2007). An integrated biomedical knowledge extraction and analysis platform: using federated search and document clustering technology. *Methods Mol Biol*, 356:293–300.

- Terra, E. and Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures. In *NAACL '03: Proc of the Conf of the North American Chapter of the ACL on Human Language Technology*, pages 165–172, Morristown, USA. ACL.
- TEWG (2003). Technical expert working group for the revision of directive 86/609/eec on the protection of animals used for experimental and other scientific purposes. Final report of the Sub-Group Ethical Review http://ec.europa.eu/environment/chemicals/lab_animals/pdf/finalreportethicalreviewprocess.pdf.
- Thut, C. J., Rountree, R. B., Hwa, M., and Kingsley, D. M. (2001). A large-scale in situ screen provides molecular evidence for the induction of eye anterior segment structures by the developing lens. *Dev Biol*, 231(1):63–76.
- Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 63–70, Morristown, NJ, USA. Association for Computational Linguistics.
- Tsuruoka, Y., McNaught, J., and Ananiadou, S. (2008). Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9(Suppl 3).
- Tuason, O., Chen, L., Liu, H., Blake, J. A., and Friedman, C. (2004). Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pacific Symposium of Biocomputing*, pages 238–49.
- Turney, P. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proc of COLING08*, pages 905–912, Manchester, UK. Coling 2008 Organizing Committee.
- Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *EMCL '01: Proc of the 12th European Conf on Machine Learning*, pages 491–502, London, UK. Springer-Verlag.
- Turney, P. D. (2003). Coherent keyphrase extraction via web mining. In *Proc of IJCAI*, pages 434–439.
- Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proc of the Int Conf on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489, Borovets, Bulgaria.
- Uschold, M. and Grüninger, M. (1996). Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155.
- Van Auker, K., Jaffery, J., Chan, J., Muller, H.-M., and Sternberg, P. (2009). Semi-automated curation of protein subcellular localization: a text mining-based approach to gene ontology (go) cellular component curation. *BMC Bioinformatics*, 10(1):228.
- Van Rijsbergen, C. J. (1979). *Information Retrieval*, 2nd edition. Dept. of Computer Science, University of Glasgow.
- Velardi, P., Navigli, R., Cucchiarelli, A., and Neri, F. (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, chapter Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies. *Frontiers in Artificial Intelligence and Applications*, ios press edition.
- Velardi, P., Navigli, R., and D’Amadio, P. (2008). Mining the web to create specialized glossaries. *IEEE Intelligent Systems*, 23(5):18–25.
- Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. In *Proceedings of the 12th Text Retrieval Conference (TREC-2003)*, pages 54–68.
- Wächter, T., Alexopoulou, D., Dietze, H., Hakenberg, J., and Schroeder, M. (2007). *Anatomy Ontologies for Bioinformatics, Principles and Practice*, volume 6, chapter Searching Biomedical Literature with Anatomy Ontologies, pages 177–194. Springer Computational Biology.
- Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. *ISMB (Supplement to Bioinformatics)*.
- Weiss, D. (2006). *Descriptive Clustering as a Method for Exploring Text Collections*. PhD thesis, Poznań University of Technology, Poznań, Poland.
- Wermter, J., Fluck, J., Strötgen, J., Geißler, S., and Hahn, U. (2005). Recognizing Noun Phrases in Biomedical Text: An Evaluation of Lab Prototypes and Commercial Chunkers. In *SMBM 2005 - Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine*, pages 25–33, Hinxton, England.
- Wermter, J. and Hahn, U. (2004). Collocation extraction based on modifiability statistics. In *Proc of COLING04*, pages 980–986, Geneva, Switzerland. COLING.
- Wermter, J. and Hahn, U. (2006). You can’t beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting*

- of the Association for Computational Linguistics, pages 785–792, Morristown, NJ, USA. Association for Computational Linguistics.
- Wernter, J., Tomanek, K., and Hahn, U. (2009). High-performance gene name normalization with geno. *Bioinformatics (Oxford, England)*.
- Westerhout, E. and Monachesi, P. (2008). Creating glossaries using pattern-based and machine learning techniques. In ELRA, editor, *Proc of LREC08*, Marrakech, Morocco.
- Wilbur, J., Smith, L., and Tanabe, L. (2007). BioCreative 2. Gene Mention Task. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 7–16.
- Winnenburg, R., Wächter, T., Plake, C., Doms, A., and Schroeder, M. (2008). Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform*, 9(6):466–478.
- Witschel, H. (2005). Using decision trees and text mining techniques for extending taxonomies. In *Proc. of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at the International Conference on Machine Learning (ICML)*, Bonn, Germany.
- Wren, J. D., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R., and Altman, R. B. (2005). Biomedical term mapping databases. *Nucleic Acids Res*, 33(Database issue):D289–93.
- Wren, J. D. and Garner, H. R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(5):426–434.
- Xu, J., Cao, Y., Li, H., and Zhao, M. (2005). Ranking definitions with supervised learning methods. In *Proc of WWW05*, pages 811–819, New York, NY, USA. ACM.
- Xu, J., Licuanan, A., and Weischedel, R. M. (2003). TREC 2003 QA at BBN: Answering definitional questions. In *Proceedings of the 12th Text Retrieval Conference (TREC-2003)*, pages 98–106.
- Yamamoto, E., Kanzaki, K., and Isahara, H. (2005). Extraction of hierarchies based on inclusion of co-occurring words with frequency information. In *IJCAI'05: Proc of the 19th Int Joint Conf on Artificial Intelligence*, pages 1166–1172, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Yang, H., Cui, H., Maslennikov, M., Qiu, L., Kan, M.-Y., and Chua, T.-S. (2003). QUALIFIER in TREC-12 QA Main Task. In *Proceedings of the 12th Text Retrieval Conference (TREC-2003)*, pages 480–488.
- Yarowsky, D. (1993). One sense per collocation. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 266–271, Morristown, NJ, USA. Association for Computational Linguistics.
- Yeh, I., Karp, P. D., Noy, N. F., and Altman, R. B. (2003). Knowledge acquisition, consistency checking and concurrency control for gene ontology (go). *Bioinformatics*, 19(2):241–248.
- Yu, H., Hripcsak, G., and Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association : JAMIA*, 9(3):262–272.
- Yu, H., Kim, W., Hatzivassiloglou, V., and Wilbur, W. J. (2007). Using medline as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *J. of Biomedical Informatics*, 40(2):150–159.
- Zhai, C., Tong, X., Milic-frayling, N., and Evans, D. A. (1997). Evaluation of syntactic phrase indexing - clarit nlp track report. In *The Fifth Text REtrieval Conference (TREC-5)*, pages 347–358.
- Zhou, L. (2007). Ontology learning: state of the art and open issues. *Inf. Technol. and Management*, 8(3):241–252.
- Zhou, W., Torvik, V. I., and Smalheiser, N. R. (2006). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–8.

Appendix

10.1 Related Work Summaries

10.1.1 Literature: Term Recognition Methods

<p>(1) <i>Automatic term recognition using contextual clues</i> (Frantzi and Ananiadou, 1997)</p> <p>(2) <i>The C-value/NC-value method of automatic recognition for multi-word terms</i> (Frantzi et al., 1998)</p> <p>(3) <i>Automatic recognition of multi-word terms: the C-value/NC-value method</i> (Frantzi et al., 2000)</p>	
<p>GOAL: Extraction of multi-word terms from corpora.</p> <p>METHOD: From the Part-Of-Speech tagged noun phrases get extracted and contextual information of candidate terms as</p> <ul style="list-style-type: none"> • $f(a)$, the total frequency of occurrence of the candidate term in the corpus • T_a, the frequency of occurrence of longer candidate terms containing a • $f(b)$, the number of these longer candidate terms, • the length of the candidate string (in number of words) <p>gets incorporated to calculate a ranking score named <i>termhood</i>, which is defined as follows:</p> $termhood(a) = f(a) - \sum_b \in T_a f(b)$ <p>The top ranked terms according to the <i>C-value</i> are used to give weights on the context. Verbals, adjectives, and nouns surrounding the potential term define the context and a weight is calculate for each such word regarding</p> <ul style="list-style-type: none"> • its total frequency of occurrence in the corpus, • its frequency as context word (of the top ranked terms), and • the number of those n-grams it appears with. <p>The <i>NC-value</i> is calculated as follows</p> $NC-value(a) = 0.8C-value(a) + 0.2 \sum_{b \in C_a} f = a(b)weight(b)$ <p>with C_a the set of distinct context words of a, $b \in C_a$, $f_a(b)$ the frequency of b as context word, $weight(b) = \frac{t(w)}{n}$ the weight of b as context word obtained from the candidate terms with top <i>C-value</i>, and $t(w)$ the. The 0.8 and 0.2 were obtained in experiments.</p>	<p>RESULTS:</p> <p>C-Value method:</p> <p>Compared to simple frequency measures and depending on the linguistic filter (extraction of phrases based on Part-Of-Speech tagged text) for the <i>C-value</i> method precision increases by 0.06 – 0.08 to 0.40 – 0.44 for those candidate terms which are nested in other terms. For terms which only occur nested precision increased by 0.31 – 0.38 to 0.50 – 0.60). Overall precision increased only 0.01 – 0.02 compared to the frequency measure and reached 0.31 – 0.38. The C-value method seems to be only little depended on the linguistic filter and the method treats term variants as separate terms.</p> <p>Context weighting factor: With additional contextual information (NC-value) the distribution of precision changes and is increased by 5% to 0.75 within the top 25% ranked candidate terms. Overall recall is not affected compared to the C-value method as candidate terms are only re-ranked.</p>

<i>You Can't Beat Frequency (Unless You Use Linguistic Knowledge) - A Qualitative Evaluation of Association Measures for Collocation and Term Extraction</i> (Wermter and Hahn, 2006)	
<p>GOAL: Testing the assumption that sophisticated statistical criteria outperform simple count of co-occurrence frequencies.</p> <p>METHOD: Automatic term recognition (ATR) methods and collocation extraction (CE) methods are tested according to four criteria for mature ATR/CE methods defined by the authors</p> <p><i>How conservative is an association measure?</i></p> <ol style="list-style-type: none"> 1. Keep the true positives (TP) in the upper ranks. 2. Keep the true negatives (TN) in the lower ranks. <p><i>How favourable is a method to undo favourable rankings?</i></p> <ol style="list-style-type: none"> 3. Demote true negatives (TN) from upper ranks 4. Promote true positives (TP) from lower ranks. <p>EVALUATION:</p> <p>8,644 manually curated collocations from a 114 million word German newspaper corpus and 31,017 from a 104 million English biomedicine corpus are used to evaluate the limited syntagmatic modifiability (LSM) for CE and limited paradigmatic modifiability (LPM) for ATR and compare them against statistical methods. LSM exploits the linguistic property, that collocations are less modifiable with with additional lexical material whereas LPM assumes that domain-specific terms are linguistically more fixed and show less distributional variation.</p> <p><i>Criterion 1:</i> t-test performs same as frequency measure; promotion of 7% from 60 to 67 for LSM and 4% from 51 to 55 for LPM in the top most segment</p> <p><i>Criterion 2:</i> against the criteria 15% of the TN get promoted to the lower part of the top segment; LSM and LPM promote approx. 30% of the TN into the upper segment.</p> <p><i>Criterion 3:</i> t-test is only marginally able to undo unfavourable rankings; LSM can demote one third of the TN in the upper segment; LPM can demote 40% of the TN from the upper segment</p> <p><i>Criterion 4:</i> promotion of 11% (CE) / 9% (ATR) of TPs from the lower segment for t-tests, which at the same time demotes TNs to the lowest segment for CE; LSM promotes 56% into lower upper segment and LSM promotes 63% in upper half segment and 24% in top most segment.</p>	<p>RESULTS: Statistical based measures e.g. t-tests, show no performance difference to the frequency of occurrence counts. LPM performs better than LSM. Criterion 2 seems to be hardest for LSM, LPM and t-test, as all methods promote TN instead of keeping them in the lower segments.</p>

Literature: Resolving Term Variations

<i>Enhancing automatic term recognition through recognition of variation</i> (Nenadić et al., 2004a)	
<p>GOAL: Evaluation of in-cooperating specific types of term variation in automatic term recognition (ATR) methods.</p> <p>METHOD: The authors compare five different types of term variation and their influence on precision and recall for ATR methods on the example of the C-Value method (Frantzi et al., 2000). The following variation types were considered:</p> <ul style="list-style-type: none"> • <i>orthographic:</i> hyphens, slashes, upper case, lower case, spelling variations and Latin/Greek spelling e.g. "amino acid" vs. "amino-acid", "NF-KB" vs. "NF-kb", "tumour" vs. "tumour", "oestrogen" vs. "estrogen" • <i>morphological:</i> inflection e.g. singular vs. plural; derivation "cell component" vs. "cellular component" • <i>lexical:</i> lexical synonyms e.g. "cancer" vs. "carcinoma" ; • <i>structural:</i> use of prepositions e.g. "clones of human" vs. "human clones"; prepositional variants e.g. "cell <u>in</u> blood" vs. "cell <u>from</u> blood"; term co-ordinations e.g. "human pancreas and liver"; • <i>acronyms and abbreviations:</i> "tuberculosis" vs. "TB" <p>The modified C – value method is linking term variations and calculated occurrence counts and termhood based on these joint term representations The original C – value method treated term variants as separate terms.</p>	<p>RESULTS: The introduction of inflection variants improved precision by approx. 25%. Acronym variations significantly improved precision by 70% when considering the most frequent terms and also improved recall up to 25%. Acronym variation detection especially lead to improvement for frequent terms, which are typically abbreviated. The in-cooperation of structural variation negatively influenced precision, because many false positives were introduced. The other variation types had only marginal influence in the ATR results.</p>

<i>Facilitating the development of controlled vocabularies for metabolomics technologies with text mining (Spasić et al., 2008)</i>	
<p>GOAL: Methodology for the rapid development of controlled vocabularies on the example of a metabolomics technology descriptions.</p> <p>METHOD: In a first step an initial controlled vocabulary (CV) gets compiled from existing resources. Using the initially selected terms, as second step the CV is refined by querying literature database PubMed for scientific abstracts and PubMed Central for full text articles, because metabolomics terminology rarely occurs in the abstract and can be found in the Materials and Methods section. The C-Value method from Frantzi et al. (2000) is used to extract the domain relevant terminology from the retrieved articles and abstracts. In step three, the CV gets discussed and evaluated by practitioners to ensure quality and completeness. As not all recognised terms focus on techniques an additional filtering step was introduced. The UMLS was used to detect outlier which contain terminology belonging to pre-defined non-relevant UMLS semantic types.</p>	<p>RESULTS: 1,600 new terms where added to the CV and manual evaluation of 100 randomly selected terms by two curators lead to a score 3.5 out of 5, where 27 and 35 out of 100 terms were definitely correct, 21 to 22 probably correct, and 7 to 37 rated as most probably wrong for terms on gas chromatography. For nuclear magnetic resonance terminology 24 to 42 were rated correct, 26 to 30 probably correct, and 17 to 41 most probably wrong or wrong.</p>

Literature: Using ontologies to extend ontologies

<i>Automatic extension of Gene Ontology with flexible identification of candidate terms (Lee et al., 2006)</i>	
<p>GOAL: Providing automatic means for predicting new concepts from the existing concepts to account for the fast accumulation of genomic data.</p> <p>METHOD: In a two step procedure the Gene Ontology (GO) is extended and the newly created concept candidates get validated. Known relationships between concepts are analysed and inherent relationships are inferred to other concepts. The terms “chemokine binding” and “C-C chemokine binding” contain the information on the hypernym relation “chemokine” → “C-C chemokine” which can now potentially become inferred to all concepts containing “chemokine” as a proper substring. Candidate terms are validated by looking for them in the biomedical literature. Since candidate terms are often of complex structure a special methods are needed to match terms (e.g. “regulation of cell differentiation”) in text. Two methods have been developed. The first method identifies the dependency structure of a sentence and verifies, if the sub-phrases of the candidate term found in the text show syntactic dependencies which correspond to their syntactic dependencies in the candidate term itself. The second method identifies the dependency structure of sentences in the abstract and is able to cross-link sentences. Again the dependency structure of the components of the candidate term are compared to the ones in the cross-linked structure.</p>	<p>RESULTS: A total of 18,964 candidate concepts were generated on the basis of 8,768 concepts of the June 2004 version of the GO. A year later it was evaluated, how many of the generated concepts are now included in the ontology. 3.5% (55/1594) of the new concepts in GO (Nov.05) could be predicted a year in advance.</p>

Literature: Named Entity Recognition

<i>Gene mention normalisation and interaction extraction with context models and sentence motifs. (Hakenberg et al., 2008)</i>	
<p>GOAL: Recognition of named entities and normalisation to a sound identifier scheme.</p> <p>METHOD: The focus of the method lies on the identification of gene mentions in text using gene specific background knowledge, such as function, location and disease annotations. To acquire knowledge dictionaries or lexicon are used which contain all known gene names for each gene. The lexica help to search for genes based on syntax. From text passages about a specific genes contextual information gets extracted and validated in gene annotation databases. An analysis of known gene names reveals variations occurring in literature not contained in the lexica. All found potential gene name mentions are referenced to a database identifier for a gene.</p>	<p>RESULTS: The method achieved an F-measure of 0.86 on the BioCreative II gene normalisation data.</p>

<i>Tagging gene and protein names in full text articles. (Tanabe and Wilbur, 2002)</i>	
<p>GOAL: Experiment on gene and protein name tagging in fulltext scientific articles</p> <p>METHOD: <i>ABGENE</i> - A Brill POS tagger is trained on 7,000 Medline/PubMed sentences using a Brill tagger with an extended lexicon now containing also entries from the UMLS SPECIALIST lexicon. The automatically generated rules from the Brill tagger are used to extract gene and protein names from biomedical abstracts. A stop word list of several thousand biomedical terms, names of amino acids, restriction enzymes, cell lines or organism names is used to filter false positives and remove non-valid "GENE" tags. False negatives are filtered by validation against a dictionary of approx 40,000 single and compound names from the former LocusLink gene name and identifier database, now included in EntrezGene¹ or share a gene context using the heuristic of a low frequent trigram is pre- or succeeded by a context word showing familiarity with gene names. In a last step documents are ranked according to their overall likelihood to contain a gene name (<i>gl</i>) and potential mentions contained in documents falling below the threshold were discarded. For fulltext articles the sentence scope was used instead.</p>	<p>RESULTS: For the evaluation of <i>ABGene</i> 2,600 sentences from PubMed Central with varying levels of <i>gl</i> were used to calculate precision and recall. The results show a maximum precision of 0.76 at 0.67 recall for one <i>gl</i> level. When strict filtering gets applied 0.62 precision at 0.60 recall is the top result for one specific <i>gl</i> level. An interpretation of these results is not given by the authors. A correlation between the <i>gl</i> level and the performance can be seen in the data.</p>

Literature: Keyphrase extraction

<i>Coherent Keyphrase Extraction via Web Mining (Turney, 2003)</i>	
<p>GOAL: Overcome incoherence of automatically extracted keyphrases</p> <p>METHOD: Candidate phrases are obtained by extracting from the corpus any sequence of one, two, or three word sequences. The following feature sets were used to train a Bayesian classifier to distinguish between keyphrase and non-keyphrase:</p> <ul style="list-style-type: none"> • Baseline feature set: tf-idf and distance, as number of words preceding the first occurrence of a term word within the document • Keyphrase frequency feature set: The baseline features and the new feature <i>keyphrase frequency</i> (the overlap between a phrase and author assigned keyphrases for all documents of the corpus except the current document) are used to classify phrases. A keyphrase is more likely to be a keyphrase if other authors used it as such. • Coherence feature set: A new feature is defined from the statistical associations between the candidate phrases and the top <i>K</i> phrases according to a classification with the baseline features. Phrases semantically related to the top phrases are preferred above others. The necessary co-occurrences are retrieved via queries to a internet search engine. • Merged feature set: This features set contains the keyphrase-frequency features set and the coherence feature set 	<p>RESULTS: In an experiment each of the four feature sets was used in training and tested on the same domain. The training set contained of 130 and test set of 500 documents selected from a corpus of Computer Science Technical Reports. Coherence features lead to a significant improvement over the baseline features. In an experiment it was confirmed that the keyphrase feature set is domain specific and its performance drops below the baseline feature performance in the interdomain evaluation.</p>

10.1.2 Literature: Abbreviation detection

<i>SaRAD: a Simple and Robust Abbreviation Dictionary (Adar, 2004)</i>	
<p>GOAL: Building an abbreviation dictionary.</p> <p>METHOD: The longforms were detected by searching for the longest common subsequence in conjunction with a set of scoring rules (Taghva and Gilbreth, 1999, see) that favours the first letter of each word of the long form. The algorithm recognises the cases, where the long form precedes the abbreviation in brackets. Morphologically similar long forms get merged if the n-grams they contain are similar. Determined based on common MeSH annotations of the associated abstracts, long forms sharing the same context get merged.</p>	<p>RESULTS: The system achieved 0.95 precision and 0.75 recall on the detection of <i>long form/abbreviation</i> pairs.</p>

¹ <http://www.ncbi.nlm.nih.gov/sites/entrez>

<i>Resolving abbreviations to their senses in MEDLINE (Gaudan et al., 2005)</i>	
<p>GOAL: Creation of a dictionary of abbreviation/sense pairs on the basis of MEDLINE abstracts and disambiguation of global abbreviations.</p> <p>METHOD: The method is based on a method by Adar (2004) described above. The authors scanned all MEDLINE and extracted 186.641 abbreviations linked to 623.441 senses represented by 5.250.259 <i>long form/abbreviation</i> pairs occurring in the ca. 3 million abstracts (years 1964 to 2004). The long forms are merged if they are similar, namely share common words between long forms. Abbreviations where no long form could be found are disambiguated based on a context model derived from Frantzi et al. (2000). With the C-value method context words are getting extracted from the abstracts. An support vector machine is trained for each sense of an abbreviation by in-cooperating all abstracts containing long forms. Before training the long forms are removed.</p>	<p>RESULTS: The system disambiguates abbreviations with a precision of 0.99, recall of 0.98 and an accuracy of 0.99</p>
<i>Building an abbreviation dictionary using a term recognition approach. (Okazaki and Ananiadou, 2006)</i>	
<p>GOAL: Recognition of acronyms in text.</p> <p>METHOD: The method is founded on the assuming that word sequences cooccurring frequently with parenthetical expressions are an expanded form of an abbreviation. The long-form recognition problem gets formalised as a term extraction problem using a modified C-value approach (Frantzi and Ananiadou, 1997)</p> <p>The authors processed all MEDLINE and extracted 886.755 candidates of acronyms and 300.954 expanded forms represented in the ca. 8 million abstracts (years 1964 to 2006).</p>	<p>RESULTS: The method achieved 0.99 precision and 0.82 – 0.95 recall on a self defined evaluation corpus.</p>
<i>ADAM: another database of abbreviations in MEDLINE (Zhou et al., 2006)</i>	
<p>GOAL: Creation of an dictionary of commonly used abbreviation, not only acronyms.</p> <p>METHOD: Five step procedure with step (1) extracting candidate abbreviations (only single word abbreviations) and surrounding text, (2) identify long forms using statistical information, (3) filter short-form/long-form pairs according to a length ration (≥ 2.5), (4) verifying that short forms are used in text separately from their long forms, and (5) grouping together morphologically similar long forms.</p>	<p>RESULTS: 97.4% and one third of the abbreviations are novel and are not contained in other databases, of which 19% of the abbreviations in ADAM are non/acronym abbreviations.</p>
<i>Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. (Yu et al., 2007)</i>	
<p>GOAL: Disambiguation of abbreviations and acronyms in full-text biomedical journals.</p> <p>METHOD: To begin a rule-based approach is used to automatically create of an dictionary of abbreviation-long form pairs from MEDLINE abstracts. Such a dictionary will contain many different long forms for each abbreviation. The following disambiguation method relies on classifiers trained on MEDLINE abstracts. Two machine learning methods, namely naïve Bayesian classification and support vector machines are trained after corresponding long forms have being normalised.</p>	<p>RESULTS: For the prediction of full forms of abbreviations in full-text articles for the two approaches precision and recall were measured. Naïve Bayesian reached for the Journal of Biological Chemistry (JBC) $p = 0.86$ and $r = 0.79$ and for the Journal of Clinical Investigation (JCI) $p = 0.90$ and $r = 0.84$. Whereas the SVM approach reached for the JBC $p = 0.89$ and $r = 0.91$ and for the JCI $p = 0.92$ and $r = 0.88$.</p>

<i>A Discriminative Alignment Model for Abbreviation Recognition (Okazaki et al., 2008)</i>	
GOAL: Development of a model for the extraction of abbreviations based on the alignment of abbreviations and their long forms. METHOD: After the creation of a 1.000 MEDLINE abstracts training corpus where abbreviations and long forms were manually aligned, a maximum entropy classifier was trained to select good combinations of features. A good combination of features consist e.g. of a set of character and bigram statements for positions in the long forms and abbreviated forms, which are correct for the manually annotated abbreviation/long form pairs.	RESULTS: The system reached a precision of 0.89 – 0.80 and recall of 0.87 – 0.98 high F-measure of 0.91 – 0.97% depending on the corpus tested.

10.1.3 Literature: Definition Generation

<i>TREC2003 QA at BBN: Answering Definitional Questions (Xu et al., 2003)</i>	
GOAL: Question Answering in TREC2003 focusing on definitional questions. METHOD: 1000 documents were retrieved containing the definiendum. Additional kernel facts were extracted from the candidate sentences to be ranked by similarity to the question targets profile using a tf-idf score. Basically the likelihood of facts sharing a context with the question target is used. The question target profile was obtained from known definitions found on web sites. 40 handcrafted rules were used to extract structured patterns that are typically used to define a term. Relation extraction techniques were used to retrieve further relational facts about the question target.	RESULTS: $F_5 = 0.55$ in the official TREC2003 task (1st rank)

<i>TREC2003 QUALIFIER in TREC-12 QA Main Task (Yang et al., 2003)</i>	
GOAL: Question Answering in TREC2003 focusing on definitional questions. METHOD: After selecting documents containing all terms of the question target. Those sentences form the positive sentence set, all other sentence the negative one. Preceding and succeeding sentences are kept. Anaphoras are replaced by the target and sentences are further ranked using two criteria: (1) sentence frequency; words of an sentence are counted in positive and negative sentence set and a score for each sentence is obtained similar to tf-idf-scores, and (2) snippets retrieved from web search engines using parts of the positive sentences are analysed for the frequency of occurrence of parts of the search target. The sentences are iteratively concatenated till the length limit for the answer is exceeded.	RESULTS: $F_5 = 0.47$ in the official TREC2003 task (2nd rank)

<i>TREC2003 Multiple-engine question answering in TextMap (Echihabi et al., 2003)</i>	
GOAL: Answering factoid questions, list questions, and definitional questions. METHOD: Answer candidates were extracted from the Web and the given corpus (TREC 2003) using sentence splitting and a maximum entropy approach to re-rank the candidates. Additionally WordNet glosses, collected biographies and descriptors for proper people as well as a set of subject-verb, object-verb, and subject-copula-object relations are used to score answer candidates. Relations are e.g. relations like "Aaron Copland composed Fanfare for the Common Man", "Aaron Copland was born in 1990"	RESULTS: The TextMap system reached a $F_5 = 0.46$ for definitional questions in TREC2003 (3rd rank)

<i>Mining Topic-Specific Concepts and Definitions on the Web (Liu et al., 2003)</i>	
GOAL: Find web pages containing definitions and find the salient concepts describing the topic of interest. METHOD: Manual collected patterns are used to find definitional sentences. HTML single structuring elements, such as headings (<h1>,<h2>,...) or emphasised text, occurring at the top of a page, are assumed to indicate a definition containing document. The hyperlinks on a page were investigated and followed to find further documents containing definitions for the term to be defined.	RESULTS: Evaluated on 28 topics the system was able to find on average 61% web pages with definitions within the top 10 results, compared to 0.18 and 0.17 precision for trivial searches with Google ² and AskJeeves ³ .

² <http://www.google.com>

³ former <http://www.AskJeeves.com>, now <http://www.ask.com>

<i>Mining on-line sources for definition knowledge (Saggion and Gaizauskas, 2004)</i>	
<p>GOAL: The definiendum, the term to be defined, provides not much valuable information for selecting the correct definitional statements from text. The authors propose other features possibly helpful for ranking candidate definitions.</p> <p>METHOD: Frequently co-occurring words are used for filtering candidate definitions assuming that co-occurring words are suitable to describe the context of the “definiendum” further. As sources for co-occurring words WordNet (Fellbaum, 1998), Britannica ⁴ and websites were used and the content was prepared using Natural Language Processing, such as tokenization, sentence splitting to create candidate definitions containing phrases.</p>	<p>RESULTS: The system answered questions from TREC 2003 and achieved $F_5 = 0.17$. When using a better text retrieval system it achieved $F_5 = 0.24$.</p>

<i>A Definitional Question Answering System Based on Phrase Extraction Using Syntactic Patterns (Han et al., 2006)</i>	
<p>GOAL: To give answers to definitional questions using shorter phrases instead of full sentences.</p> <p>METHOD: Keyword searches are used to retrieve text passages, followed by sentence splitting and filtering to obtain all sentences containing the words to be defined. With the help of syntactic patterns, noun phrases, verbal phrases, pronoun phrases and participle phrases get retrieved as initial answer candidates. The answer candidates are ranked based on several criteria, namely (a) the redundancy of the term to be defined, (b) the importance of words based measured by their local term statistics, defined as $Loc(C) = (\sum_{t_i \in C} \frac{sf_i}{maxsf}) / C$, where sf_i is the number of sentences containing the term and $maxsf$ describes the maximal number of sentences containing one of the terms, (c) the conditional probability of a statement C, under the conditions that each of the external definitions obtained from a dictionary or encyclopedia is also a valid answer to the question, and finally (d) the probability of vocabulary used in definition in comparison to general text.</p>	<p>RESULTS: The system was evaluated on 50 TREC'2003 and 64 TREC'2004 topics and it was shown, that short answer phrases lead to better overall performance. Drawback is the generally low $F_1 = 0.16$, with a $Recall = 0.34$ and $Precision = 0.09$.</p>

<i>Automated detection and annotation of term definitions in German text corpora (Storrer and Wellinghoff, 2006)</i>	
<p>METHOD: In a first step, definitions are extracted from German text using classical patterns. In a second step the possibility to extract hypernym, hyponym and holonym relations from the candidate definitions is evaluated in feasibility study. The authors aim to extract the definitions in the form of semantic relations between the head nouns of the definiendum and the head nouns of the definiens. The study focuses on the defining verb, the definitor.</p>	<p>RESULTS: Most common definitor are forms of “to be”, like “is a” or “are” (“ist ein” und “sind” in German). Especially these forms of “to be” are used equally in definitions and other text. The evaluation lead to 0.31 precision at 0.83 recall based on 80 statements containing forms of “to be”. Overall 0.34 precision at 0.70 recall where obtained for 19 different verbals playing the role of a definitor.</p>

Literature on definition extraction using grammars or parsers

<i>DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text (Klavans and Muresan, 2000)</i>	
<p>METHOD: Patterns for definitions were extracted from the cardio corpus⁵ using a finite state grammar to model rules to extract patterns to be used for the extraction of definitions.</p>	<p>RESULTS: DEFINDER was evaluated in Klavans and Muresan (2001) where the system identified 40 out of 53 definitions obtaining 0.87 precision and 0.75 recall. In a empirical evaluation the author state that especially the usefulness of DEFINDER retrieved definitions and their readability outperforms those definitions found in the UMLS or Online Medical Dictionary.</p>

⁴ <http://www.britannica.co.uk>

⁵ <http://www.cardio.com/articles.html>

<i>Towards the Automatic Extraction of Definitions in Slavic (Przepiórkowski et al., 2007)</i>			
METHOD: An experiment was performed for the semiautomatic glossary construction from texts in Bulgarian, Czech, and Polish using grammar evaluated over morphosyntactically-annotated documents.	RESULTS: For each language a test corpus of 150 – 200 manually annotated definitions was used to measure the F-measure. Results are generally described as not satisfactory.		
		precision	recall F_2
	Bulgarian	0.23	0.09 0.11
	Czech	0.22	0.46 0.34
	Polish	0.23	0.32 0.28
Results for Bulgarian are especially very low, because 0.36 of definitions in the corpus are spread over multiple sentences which makes the extraction more difficult.			

Literature on definition extraction using machine learning

<i>Unsupervised Learning of Soft Patterns from On-line News (Cui et al., 2004)</i>	
METHOD: Introduction of soft pattern, namely learnt vectors of words and syntactic classes to overcome the inflexibility of manual created patterns which are matched slot by slot.	RESULTS: $Recall = 0.60$, $Precision = 0.22$, and $F_1 = 0.53$ evaluated on 50 TREC2003 questions. Selecting sentences containing definition with $Precision = 0.33$ within the top 10 ranked sentences.

<i>Creating glossaries using pattern-based and machine learning techniques (Westerhout and Monachesi, 2008)</i>																	
In web pages, definitions are stated in many ways and this work investigated this and distinguished between five types how terms in natural language text get defined: to be: "Liver is an organ which is present in vertebrates and some other animals." verb: "The primary spinal tumour affects the spinal cord cell and nerve roots." punctuation: "Liver: organ present in vertebrates and some other animals" layout: Liver "Organ present in vertebrates and some other animals" pronoun: "Liver. This is an organ present in vertebrates and some other animals"	<p>Evaluated on 330 manually annotated definitions, the authors quantified the use of the different types:</p> <table> <tr> <th>Type</th><th>Number (percentage)</th></tr> <tr> <td>to be</td><td>84 (25.5%)</td></tr> <tr> <td>verb</td><td>99 (30%)</td></tr> <tr> <td>punctuation</td><td>46 (13.9%)</td></tr> <tr> <td>pronoun</td><td>46 (13.9%)</td></tr> <tr> <td>layout</td><td>7 (2.1%)</td></tr> <tr> <td>other patterns</td><td>48 (14.5%)</td></tr> <tr> <td>definition contexts</td><td>330</td></tr> </table> <p>Based on the 330 definitions an newly annotated 150 definition they evaluated the performance of the grammar based approach for finding definitions. to be types could be found nearly as correct as the learning examples (F_2 0.51 \rightarrow 0.43). Results for the verb type on the other side were in the test corpus much lower than in the training data (F_2 0.62 \rightarrow 0.35). The results for the types punctuation and pronoun were very low. After filtering using a machine learning approach the results for to be and punctuation type definitions could be increased to $F_2 = 0.71$ and $F_2 = 0.40$</p>	Type	Number (percentage)	to be	84 (25.5%)	verb	99 (30%)	punctuation	46 (13.9%)	pronoun	46 (13.9%)	layout	7 (2.1%)	other patterns	48 (14.5%)	definition contexts	330
Type	Number (percentage)																
to be	84 (25.5%)																
verb	99 (30%)																
punctuation	46 (13.9%)																
pronoun	46 (13.9%)																
layout	7 (2.1%)																
other patterns	48 (14.5%)																
definition contexts	330																

<i>Definition Extraction using a Sequential Combination of Baseline Grammars and Machine Learning (Degórski et al., 2008)</i>	
<p>METHOD: Definitions get extracted using an automatically trained machine learning classifier without the need to construct sophisticated manual grammars. A corpus with approx. 300K tokens and over 550 definitions was used to automatically Part-Of-Speech tagged and definitions got manually annotated. From the total of 558 definition, 386 definitions were used in the training of the classifier and 172 definitions as test corpus.</p>	<p>RESULTS: The authors evaluated the performance of different machine learning classifiers and grammar based methods. Additionally the performance gain for applying the grammar prior the classifier was measured. The pure grammar-based approach reached a maximal F-measure $F_1 = 0.28 (F_5 = 0.44)$ with $Recall = 0.59$ at $Precision = 0.19$. Machine learning classifiers reached at most $F_1 = 0.27 (F_5 = 0.30)$. It was observed that recall improved when restricting the ratio of positive to negative samples to e.g. $\frac{1}{5}$ by randomly choosing the negative samples from the set of all negative samples. The influence of sampling was observed to rarely exceed 0.5% for both precision and recall. By combining the methods, the F-measure can further be improved. The grammar used in the experiment rejects 12% of the sentences and improves this way precision significantly and only marginally reduces recall, e.g. for one combination $F_1/F_2/F_5$ improves from 0.18/0.24/0.35 to 0.30/0.36/0.46. True improvement only is achieved when “machine learning algorithms are supported by some – relatively trivial – a priory linguistic knowledge”.</p>

<i>Mining the Web to Create Specialized Glossaries (Velardi et al., 2008)</i>	
<p>METHOD: The system extracts definitions from the web search results and additionally uses Google’s <i>define</i> to generate definition candidate. Definitions are ranked according to a so called stylistic filter which accepts only definitions where the genus and the differentia is present. The filter is implemented as machine learning classifier based on a training set of > 100 positive and > 50 negative definition sentences from the domains “arts”, “tourism”, “computer networks”, and > 1000 positive and negative sentences from the domain “enterprise interoperability”.</p>	<p>RESULTS: The evaluation of 359 predicted definitions from web documents lead to an $F_1 = 0.86$.</p>

10.1.4 Literature: Taxonomy Induction

Literature on Methods using Syntactic Patterns

<i>Automatic acquisition of Hyponyms from large text corpora (Hearst, 1992)</i>	
<p>GOAL: Automatic acquisition of Hyponyms avoiding pre-encoded knowledge and applicability of the method over a wide range of text.</p> <p>METHOD: Hearst compiled a set of lexico-syntactic patterns usually used to describe subsumption (mainly hypernymy) in text. Examples are: <i>A is a B</i> or <i>B such as A</i>. With these patterns one can infer e.g. from the text fragment “organelles such as mitochondria”, that mitochondria are organelles.</p>	<p>RESULTS: The authors analysed corpora for existence of such patterns:</p> <ul style="list-style-type: none"> • Grolier’s American Academic Encyclopedia (8.6M words) and 7067 sentences containing “such as” of which 152 were following the strict criteria, that hyponym and hypernym are unmodified. • New York Times news corpus (20M words) and 3178 sentences containing “such as” with 42 relations according to the strict criteria (see above) <p><i>Hearst-patterns have high precision, but a low recall, since many relationships are not made explicit in text.</i></p>

<p><i>A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition (Faure and N'edellec, 1998)</i></p> <p><i>Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM (Faure and N'edellec, 1999)</i></p>	
<p>GOAL: Learning subcategorisation frames (SF) of verbs and ontologies from syntactic parsing of technical texts in natural language. An example for such an instantiated SF is <to travel> <subject: [father, neighbour, friend]> <by: [car, train]>.</p> <p>METHOD: By syntactic parsing all interpretations of the parsed sentences are being obtained and used as input for the ASIUM system. As interpretation of concepts nouns are grouped if they share at least two SF (verb+preposition/syntactic role). The meaning of a concept is defined by set of SFs assigned to it. Bottom-up breadth-first clustering gets applied to aggregate concepts of the ontology level by level. Concepts and labels get validated by an expert. Only validated concepts can participate in the construction of new concepts.</p>	<p>RESULTS: The goal of learning the full concept hierarchy cannot be achieved fully automatically. User input is required for validating concepts and concept labels.</p>
<p><i>Learning syntactic patterns for automatic hypernym discovery (Snow et al., 2004)</i></p>	
<p>GOAL: Development of an algorithm for automatically learning hypernyms, as generalized approach for previous methods relying on regular expression patterns.</p> <p>METHOD: Machine learning is used to learn lexico-syntactic patterns which are combined in a hypernym classifier. A number of steps are performed for training:</p> <ol style="list-style-type: none"> Collect noun pairs from corpora (examples of hypernyms) For each noun pair, collect sentences containing the nouns Parse and extract patterns from the parse tree Train a hypernym classifier <p>Noun pairs can now be tested for hypernymy using the trained machine learning classifier.</p>	<p>RESULTS: Compared to Hearst Patterns the machine learning classifier reaches an improvement of 132% average F-measure. The evaluation against WordNet lead to an improvement of 54% in F-measure for the best classifier. This classifier takes coordinate terms into account. Coordinate terms are terms that share the same hypernym. This allows to infer hypernyms if the coordinate terms is known. Coordinate terms can be found using coordination patterns, such as "X,Y, and Z" or distributional similarity measures (Pereira et al., 1993). Additionally it was found that the ratio between hypernym and non-hyponym word pairs was approx. 1:50, as well as all patterns originally suggested by Hearst (1992) where also identified by the proposed generalized method. The authors found a maximal F-measure of 0.14 (Hearst Patterns), 0.23 (WordNet), 0.27 (TREC hypernyms), 0.33 (TREC hypernyms + coordinate terms), and 0.36 (TREC + Wikipedia hypernyms + coordinate terms).</p>
<p><i>Semantic taxonomy induction from heterogeneous evidence (Snow et al., 2006)</i></p>	
<p>GOAL: Incorporation of evidence from multiple classifiers over heterogeneous relationships to optimise the structure of an taxonomy.</p> <p>METHOD: A probabilistic model was created which defines the "conditional probability for a set of relational evidence given a taxonomy". Taxonomy learning gets redefined as a local search problem of finding the taxonomy maximising this conditional probability. Using a best-first search algorithm in each iteration of the search algorithm a new taxonomy is created from the union of the previous taxonomy and a new set of relations. This new set is the set of relations implied by the relation which maximises conditional probability of the taxonomy given a set of evidences (syntactic patterns) for all the single relations.</p>	<p>RESULTS: The approach was used to extend WordNet version 2.1 and reports approx. 0.20 recall at 0.58 precision.</p>

Literature on Methods using Similarity Measures

<i>Automatic construction of a hypernym-labelled noun hierarchy from text (Caraballo, 1999)</i>	
<p>GOAL: Building of a labeled noun hierarchy based on text. Nouns are clustered into a hierarchy using data on conjunctions of noun phrases like and appositives (e.g. "Dresden, a city in Germany").</p> <p>METHOD: (1) Conjunctions and appositives were collected from the Wall Street Journal corpus. For each noun a vector is created containing the count of how often an other noun occurred together with it in a conjunction or appositive. Cosine similarity was used to measure the similarity between a pair of vectors. The similarity of two groups of words A and B is computed as follows:</p> $\text{sim}(A, B) = \frac{\sum_{v,w} \cos(v, w)}{ A \cdot B }, v \in A \text{ and } w \in B$ <p>(2) Hypernyms get extracted as described by Hearst (1992) using manual syntactic patterns.</p>	<p>RESULTS: The algorithm produces correct hyponyms in 33% of all cases. The evaluation is based on a sample of 10 nodes each dominating at least 20 nouns. The total tree contained 20,014 nouns which have been structured by 654 nodes. Up to three hypernyms were listed as "best" hypernyms for each node. Three human judges had to assess for each noun whether the hypernyms assigned to the corresponding nodes are correct. For 60% of the tested nouns at least one judge judged one hypernym as correct.</p>
<i>Deriving Concept Hierarchies from Text (Sanderson and Croft, 1999)</i>	
<p>GOAL: "Deriving a concept hierarchy from a set of documents without the use of training data or standard clustering techniques."</p> <p>METHOD: Phrases are getting hierarchically organised; subsumption is tested using co-occurrence data for words and phrases. Iff $P(x y) \geq 0.8$, $P(y x) < 0.8$ (<i>subsumption criterion</i>) holds, the authors assumed a subsumption relation $x \rightarrow y$.</p>	<p>RESULTS: 48% of the term pairs fulfilling the <i>subsumption criterion</i> where judged of having an "aspect of" or "type of" relationship for a hierarchy (similar to precision). For random term pairs 28% were observed (possible baseline).</p>
<i>Collaborative creation of communal hierarchical taxonomies in social tagging systems. (Heymann and Garcia-Molina, 2006)</i>	
<p>GOAL: Extracting directed taxonomic relations from text.</p> <p>METHOD: In this method two terms are linked if the cosine similarity of their document vectors is above a threshold. The term, which is more central in the whole graph, becomes the parent, the other the child.</p>	<p><i>An evaluation of the algorithm will be given in Chapter 5 (Taxonomy Generation)</i></p>
<i>Using Decision Trees and Text Mining Techniques for Extending Taxonomies (Witschel, 2005)</i>	
<p>GOAL: Development of a semi-automatic procedure for extending lexical taxonomies using ATR methods.</p> <p>METHOD: The method identifies noun phrases with a pattern based approach using Part-Of-Speech tags, selects candidate terms based on frequency and locates them in a hierarchy by utilizing co-occurrence features from large corpora.</p>	<p>RESULTS: Regarding taxonomy construction the results are poor in terms of learning accuracy and do not show a significant difference between the best run and the baseline, where all concepts are simply located as hyponym of the root node. Even though a huge learning corpus (ca. 5 GB) was used the classification data was sparse. Only for 60% of the chosen example, a minimum of 10 similar words could be found. In a qualitative evaluation the classification of top level concepts was tested with on an artificial ontology containing two overlapping subtrees from GermanNet. The classification performance was measured pairwise for the root nodes of each sub-tree. Recall was observed to be much lower than precision and the ration of size (number of concepts) between the compared subtrees seemed to influence the algorithm significantly and overall F-measure drops below the given baseline, defined as a classification of all nodes under the bigger sub-tree. The best accuracies observed were between 11% and 14%.</p>

<i>Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis</i> (Cimiano et al., 2005)	
<p>GOAL: Development of an method to obtain an taxonomy from natural language text.</p> <p>METHOD: The authors describe a method, that clusters the candidate terms into a hierarchy by identifying evidence to hypernymy relations among them from literature. At first pairs of verbs and subject complements, object complements or preposition phrase complements are extracted from text using linguistic parsers, e.g. The university created a new institute. Where: $create_{subj}(university)$ and $create_{obj}(institute)$ simply capturing that an institute can be created and that an university can create things. Using formal concept analysis a lattice is derived which was converted in a partial order constituting a concepts hierarchy.</p>	<p>RESULTS: On two domain examples <i>Tourism</i> and <i>Finance</i> the FCA approach was evaluated and compared with KMeans and hierarchical clustering. With $F_{Tourism} = 0.41$ and $F_{Finance} = 0.33$, FCA outperformed all clustering methods in terms of F-measure. This is due to higher recall values mainly. A drawback of FCA is the exponential time complexity of $O(2^n)$ compared to only $O(n^2)$ or $O(n^2 \log n)$ for KMeans and agglomerative clustering methods.</p>
<i>Taxonomy Learning using Term Specificity and Similarity</i> (Ryu and Choi, 2006)	
<p>GOAL: Analysis of features for specificity and similarity in previous methods and selection of optimal features to be used for taxonomy learning.</p> <p>METHOD: Term specificity is a necessary condition for taxonomy learning, because specific terms tend to be locate in low level of a domain taxonomy. Term similarity is a necessary condition in taxonomy learning, because similar terms group close together in a taxonomy. Therefor it is highly probable that that term t_1 is an ancestor of t_2 in a taxonomy T_D, if both are semantically similar and t_2 is more specific than t_1 in the domain D.</p> <p><i>Features for specificity of terms:</i></p> <ul style="list-style-type: none"> • $Spec_{adj}$ – term t (a noun) is specific, if there are few adjectives modifying it (Caraballo, 1999; Ryu and Choi, 2005) • $Spec_{varg}$ – Verb-argument distribution is based on the co-occurrence of terms with special verbs. A term is more specific, if it co-occurs frequently with the same verbs. E.g. "protein" and "increase", "activate", "inhibits", "binds", etc. (Cimiano et al., 2005) • $Spec_{coldoc}$ – Conditional probability of term co-occurrence regards a term t_a to subsume t_b, if $P(t_a t_b) > P(t_b, t_a)$. Hence t_b is more specific then t_a. • $Spec_{in}$ – Inside-word information is used to measure specificity for multiword terms. Indicates what component word which is highly associated with a term contributes specificity to the term. • $Spec_{in/adj}$ – harmonised similarity from $Spec_{in}$ and $Spec_{adj}$ to regard both inside and outside information. <p><i>Features for similarity of terms:</i></p> <ul style="list-style-type: none"> • If terms co-occur in similar documents, they are similar (Sanderson and Croft, 1999). • If vectors of adjective patterns of terms are similar, the terms are similar (Yamamoto et al., 2005) • If vectors of verb-argument dependencies are similar, the terms are similar (Cimiano et al., 2005). 	<p>RESULTS: Ryu and Choi compared four taxonomy learning methods and reported recall below 0.50 and a precision of 0.50 or lower. It was tested whether the assumption holds, that in a valid parent-child relationship the specificity of the parent is lower than the specificity of the child. While $Spec_{adj}$ showed the highest precision, recall was very low as usually there exist few modifications of nouns by adjectives. Regarding <i>similarity</i> it was observed, that taxonomy based similarity ratings are closest to human similarity ratings (correlation coefficient of 0.85).</p>

10.2 Figures and Tables

Supplementary Material from Wächter and Schroeder (2010)

Due to the size of the tables it is not possible to layout them in the appendix of this thesis. The tables 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, and 10.7 can be accessed under <http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/> and after publication as supplementary tables to Wächter and Schroeder (2010) at the publishers website.

Table 10.1. Listing of 1,000 randomly selected MeSH terms for term generation
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup3.xls>

Table 10.2. Generated terms for 1,000 MESH terms and mapping to GO, MeSH, OBO, and UMLS.
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup4.xls>

Table 10.3. Listing of 500 randomly selected GO terms for definition generation.
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup5.xls>

Table 10.4. Listing of 500 randomly selected MeSH terms for definition generation.
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup6.xls>

Table 10.5. Evaluation TREC 2003: questions and manual curation of automatically generated answers
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup7.xls>

Table 10.6. Evaluation GO definitions: listing of manual curation for top 10 generated definitions.
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup8.xls>

Table 10.7. Evaluation MeSH definitions: listing of manual curation for top 10 generated definitions.
<http://www.biotec.tu-dresden.de/~waechter/DOG4DAG/waechter-sup9.xls>

<i>Experiment</i>	GO	biological process	cellular component	molecular function
Total	28814	17541	2614	8659
Terms with definition	91.1%	99.6%	100%	97.9%
Words in definition	24.3	27.1	28.8	16.9
	(± 15.3)	(± 15)	(± 20)	(± 10)
Terms in definition	2.4 (± 2.3)	2.8 (± 1.1)	3.9 (± 2.8)	1.3 (± 1.4)
≤ 1 term in definition	88.0%	90.0%	96.2%	81.3%
≤ 1 parent in definition	15.8%	17.7%	49.8%	1.4%
≤ 1 ancestor in definition	54.1%	74.4%	85.5%	2.6%

Table 10.8. Proportion of terms in GO parts *biological_process*, *cellular_component*, and *molecular_function* containing parent terms, ancestor terms or other existing terms in their definitions.

<i>Experiment</i>	<i>MeSH</i>		<i>Anatomy</i>		<i>Biological Sciences</i>		<i>Chemicals and Drugs</i>		<i>Diseases</i>		<i>Geographicals</i>		<i>Health Care</i>		<i>Information Sciences</i>		<i>Named Groups</i>		<i>Natural Sciences</i>		<i>Organisms</i>		<i>Psychiatry and Psychology</i>		<i>Techniques and Equipment</i>		<i>Technology, Industry, Agriculture</i>	
Total	29348		1669		1995		8832		4420		381		1621		524		201		1517		3580		1093		2698		817	
Terms with definition	96.0%		92.6%		98.5%		95.7%		97.6%		54.6%		97.0%		91.4%		95.5%		96.6%		99.0%		99.3%		95.6%		94.1%	
Words in definition	30.2		32.2		29.7		30.2		35.8		52.1		26.5		31.2		19.2		28.7		25.3		27.8		30.3		28.1	
Terms in definition	±19.3		±18.7		±19.7		±17.4		±23.4		±34.8		±18.5		±20.9		±12.1		±18.6		±14.1		±20.2		±19.4		±17.1	
≤ 1 term in definition	5.7 ±4.1		6.3 ±3.6		5.1 ±4.1		5.9 ±4.2		7.6 ±5.2		7.1 ±3.5		4.2 ±3.3		4.3 ±3.4		3.4 ±2.5		4.5 ±3.4		5.7 ±3.0		4.2 ±4.2		4.9 ±3.6		5.0 ±3.3	
≤ 1 ancestor in definition	97.2%		99.6%		95.7%		96.7%		99.3%		100.0%		94.8%		94.6%		96.4%		95.6%		99.8%		90.6%		96.7%		96.9%	
≤ 1 parent in definition	56.2%		80.6%		38.7%		52.5%		51.9%		73.6%		49.7%		44.9%		54.7%		46.3%		92.2%		40.3%		39.4%		64.8%	
≤ 1 parent in definition	36.6%		62.3%		28.1%		31.7%		26.7%		66.8%		28.7%		25.5%		26.6%		29.7%		72.4%		27.6%		23.6%		49.0%	

Table 10.9. Proportion of term mentions in term definitions analysed for the MeSH trees *Anatomy, Biological Sciences, Chemicals and Drugs, Diseases, Geographicals, Health Care, Information Sciences, Named Groups, Natural Sciences, Organisms, Psychiatry and Psychology, Techniques and Equipment, and Technology, Industry, Agriculture* containing parent terms, ancestor terms or other existing terms in their definitions.

Question	Positions of valid definitions
Qid_1901__Aaron_Copland	2 5 6 24
Qid_1905__a_golden_parachute	1 2 3 4 5 6 7 8 9 13 15
Qid_1907__Alberto_Tomba	12
Qid_1917__Bausch_&_Lomb	2
Qid_1933__Vlad_the_Impaler	1 4
Qid_1955__Akbar_the_Great	
Qid_1957__fractals	3
Qid_1964__Allen_Iverson	1 9 12 19
Qid_1972__Abraham_in_the_Old_Testament	
Qid_1987__ETA_in_Spain	1 8
Qid_2006__Aga_Khan	1 2
Qid_2008__the_vagus_nerve	1 3
Qid_2024__Andrea_Boccelli	1
Qid_2041__Iqra	12 22
Qid_2042__Abu_Sayaf	1 14
Qid_2060__Albert_Ghiorso	2
Qid_2082__Anthony_Blunt	
Qid_2095__TB	3 7 11
Qid_2112__Antonia_Coello_Novello	1 5
Qid_2125__Charles_Lindberg	1 7 12
Qid_2130__Ben_Hur	1 13 19
Qid_2146__Bill_Bradley	1 2 26
Qid_2148__Ph_in_biology	
Qid_2150__El_Shaddai	5 24
Qid_2158__the_Hague	1 29 31
Qid_2174__Alexander_Hamilton	53
Qid_2177__Angela_Davis	2 3 13 41 54
Qid_2201__Bollywood	3 10 24
Qid_2203__a_quasar	5 7 9 20
Qid_2208__Al_Sharpton	2 3
Qid_2222__Friends_of_the_Earth	3 14 15 16 18 20 28
Qid_2224__Andrew_Carnegie	41
Qid_2229__Freddie_Mac	1 8 12 16 40
Qid_2234__Althea_Gibson	5
Qid_2258__feng_shui	4 5 8 11
Qid_2267__Alexander_Pope	1 4
Qid_2274__Alice_Rivlin	3 5 7
Qid_2304__Niels_Bohr	1 28
Qid_2321__Restorative_Justice	1 2 3 4 5 6 7 8 9 10
Qid_2322__Absalom	27 33 57
Qid_2324__Nostradamus	12 33 40 83
Qid_2327__Ari_Fleischer	16
Qid_2332__Machiavelli	2 6 7
Qid_2348__the_medical_condition_shingles	
Qid_2349__Anwar_Sadat	1 11 20
Qid_2366__Schadenfreude	2 5 6
Qid_2369__Annie_Oakley	1 2
Qid_2372__Destiny's_Child	3 4 8
Qid_2373__Alger_Hiss	6 17
Qid_2385__the_Kama_Sutra.txt	1

Table 10.10. Overview over correctly generated definition by DOG4DAG from the TREC2003 definitional question answering task. Positions of retrieval are shown for correctly generated TREC2003 answers.

Fig. 10.1. Listing of all 152 terms in branch 3Rs in Human Toxicity Testing of the Go3R ontology with semi-automatically created definitions. For 65 (43%) of the human toxicity testing methods a definition could be found, for 41(27%) the top ranked generated definition has been chosen. The comment lists for each term with a definition the retrieval rank. For presentation, synonyms for the terms have been removed.

format-version: 1.2 date: 14:0rep15:2010 02:59 saved-by: waechter auto-generated-by: OBO-Edit 2.1-beta4	comment: Generated 1st and 10th	name: HSP Reporter Gene Assay
[Term] id: 000125526 name: In Vitro Skin Sensitisation Testing comment: "No description available."	[Term] id: 000125703 name: MTT Assay def: "MTT assay is an index of cell viability and cell growth, which is based on the ability of viable cells to reduce MTT from a yellow water-soluble dye to a purple-insoluble formazan product." [URL ...] comment: Generated 2nd	[Term] id: 000125732 name: HSP RT-PCR
[Term] id: 000125527 name: Direct Peptide Reactivity Assay	[Term] id: 000125704 name: XTT Assay def: "XTT assay is a colorimetric method of quantifying fungal growth by measuring the metabolism of XTT by fungal mitochondrial dehydrogenases. XTT assay is a non-radioactive alternative for the [51 Cr] release cytotoxicity assay." [URL ...] comment: Generated 1st and 3rd	[Term] id: 000125734 name: iNOS Western Blot
[Term] id: 000125528 name: Myeloid U937 Skin Sensitisation Test	[Term] id: 000125706 name: Stress Response Testing	[Term] id: 000125735 name: iNOS Reporter Gene Assay
[Term] id: 000125529 name: Human Cell Line Activation Test def: "human Cell Line Activation Test is an in-vitro skin sensitization method based on the enhancement of CD86 and/or CD54 in THP-1 cells." [URL ...] comment: Generated 1st	[Term] id: 000125707 name: ROS Assay def: "[Both the MTS assay and the] ROS assay are reliable assays to determine toxic effects of silver nanoparticles in this cell line." [URL ...] comment: Generated 2nd	[Term] id: 000125736 name: RNS/NO Assay
[Term] id: 000125699 name: BrdU Assay def: "[Thus, the] BrdU assay is an alternative non-radioactive assay for the determination of cell proliferation." [URL ...] comment: Generated 5th	[Term] id: 000125708 name: Modified Comet Assay	[Term] id: 000125765 name: Coagulation Assay
[Term] id: 000125700 name: WST-1 Assay def: "WST-1 assay is a colorimetric assay that tests proliferation and viability of cells based on the ability of viable cell mitochondrial dehydrogenases to cleave a tetrazolium salt (WST-1) substrate." [URL ...] comment: Generated 2nd	[Term] id: 000125709 name: GSH Depletion Assay def: "Glutathione (GSH) depletion is an early hallmark observed in apoptosis, and we have demonstrated that GSH efflux during death receptor-mediated apoptosis occurs via a GSH transporter." [URL ...] comment: Generated 3rd	[Term] id: 000125766 name: DCF-DA ROS Assay
[Term] id: 000125701 name: Alamar Blue Assay def: "Alamar Blue assay is a quantitative measurement of the proliferation of human and animal cell lines which incorporates a fluorometric/colorimetric growth indicator based on detection of metabolic activity." [URL ...] comment: Generated 1st	[Term] id: 000125710 name: SOD Assay def: "SOD Assay is a convenient colorimetric SOD inhibition activity assay." [URL ...] comment: Generated 1st	[Term] id: 000125767 name: D-dimer ELISA
[Term] id: 000125702 name: MTS Assay def: "MTS assay is a colorimetric assay based upon the ability of viable cells to convert MT to formazan; the quantity of formazan product, as measured by the 490 nm absorbance, is directly proportional to the number of viable cells in culture. MTS assay is a direct measurement of cell viability, from which virus-mediated cytotoxicity was quantitated." [URL ...]	[Term] id: 000125728 name: Nrf2 Reporter Gene Assay	[Term] id: 000125768 name: Thrombin Activity Assay
	[Term] id: 000125729 name: HSP Microarray	[Term] id: 000125775 name: Neuronal Differentiation Assay
	[Term] id: 000125730 name: HSP Western Blot	[Term] id: 000125781 name: In Vitro Barrier Tests subset: Tox terme
	[Term] id: 000125731	[Term] id: 000125792 name: Flow Cytometry Assay def: "[In conclusion, the] flow cytometry assay is a quantitative method for the detection of cell-mediated cytotoxicity and does not use radioactive labeling." [URL ...] comment: Generated 2nd
		[Term] id: 000125793 name: Calcein Assay def: "Calcein Assay is a proprietary indirect inhibitory whole-cell assay that provides information on any interaction between the ABC transporter (MDR1/P-gp (ABCB1) or MRP1 (ABCC1 ...)" [URL ...] comment: Generated 1st
		[Term] id: 000125794 name: Digoxin Bidirectional Transport Interaction
		[Term] id: 000125795

XXX 10 Appendix

name: Hoechst Assay	comment: Generated 1st	[Term]
[Term]	[Term]	id: 000125886
id: 000125796	id: 000125863	name: Protein Kinase Phosphorylation Assay
name: Membrane Based Transporter Assay	name: Single Cell Gel Electrophoresis	
[Term]	def: "Single cell gel electrophoresis is a sensitive technique for monitoring the damage to cellular DNA by various genotoxic agents like radiation." [URL ...]	[Term]
id: 000125798	comment: Generated 1st	id: 000125919
name: In Vitro hERG Blocking Potency Assay		name: Neutral Comet Assay
[Term]		def: "neutral comet assay is a useful method allowing direct visualization of DNA damage, mainly DSBs." [URL ...]
id: 000125803		comment: Generated 2nd
name: Rubidium Flux Assay	[Term]	
[Term]	id: 000125864	[Term]
id: 000125804	name: Cytokinesis Blocked Micronucleus Assay	id: 000125920
name: Competitive Radioligand Binding Assay	def: "cytokinesis-blocked micronucleus assay is a short-term muta-genesis test which offers an easier and less tedious alternative to metaphase chromosome analysis, with the advantage that exposure to both clastogens and aneugens may be detected." [URL ...]	name: Alkaline Comet Assay
[Term]	comment: Generated 1st	def: "alkaline Comet assay is a very useful method for studying genotoxicity in cells exposed in vitro or in vivo to a variety of physical and chemical agents (Tice et al., 2000)." [URL ...]
id: 000125805		comment: Generated 1st
name: In Vitro Electrophysiology Measurement	[Term]	
[Term]	id: 000125868	[Term]
id: 000125809	name: TUNEL Assay	id: 000125932
name: Fluorescent Imaging Plate Reader Membrane Potential Method		name: 3D Model Liver Tissue
[Term]	[Term]	
id: 000125810	id: 000125874	[Term]
name: Patch-Clamp Technique	name: Luciferin-Luciferase Assay	id: 000125933
[Term]		name: Liver Tissue Spheroid
id: 000125815	[Term]	
name: Cell Viability Assessment	id: 000125875	[Term]
[Term]	name: Vialight Assay	id: 000125953
id: 000125816	[Term]	name: Flow Through Diffusion Cell
name: Lactate Dehydrogenase Assay	id: 000125877	[Term]
def: "lactate dehydrogenase assay is a means of measuring either the number of cells via total cytoplasmic lactate dehydrogenase (LDH) or membrane integrity as a function of the amount of cytoplasmic LDH released into the medium." [URL ...]	name: Colony Formation Assay	id: 00042719
comment: Generated 1st	def: "colony formation assay is a stringent assay for long-term cell proliferation that requires not only cell division but also subsequent cell survival and adhesion to the plate." [URL ...]	name: SkinEthic
[Term]	comment: Generated 1st	def: "SkinEthic is an important worldwide player in the production and commercialisation of human epidermal and epithelial tissues (including epidermis, dermis, corneal, oral, gingival, oesophageal epithelium, alveolar and vaginal mucosa) for in vitro test applications across many industries." [URL ...]
id: 000125817	[Term]	comment: Generated 1st
name: Glucuronidase Assay	id: 000125879	
[Term]	name: Colorimetric Assays	[Term]
id: 000125839	def: "colorimetric assay is a quantitative chemical analysis measuring color intensity produced by reacting a sample with a reactant as a proxy for the amount of the assayed material in a sample." [URL ...]	id: 00042722
name: Toxilight	comment: Generated 1st	name: In Vitro Human Skin Assays
[Term]		
id: 000125840	[Term]	[Term]
name: Luciferase Assay	id: 000125883	id: 00042724
def: "Luciferase is a generic term for the class of oxidative enzymes used in bioluminescence and is distinct from a photoprotein. Luciferase assay is a fast, easy and sensitive assay (detection of a light signal)." [URL ...]	name: Trypan Blue Exclusion Assay	name: ICE Test
comment: Generated 1st and 12th	def: "trypan blue exclusion assay is a standard assay of cell viability." [URL ...]	def: "Isolated Chicken Eye (ICE) Test Method for Identifying Ocular Corrosives." [URL ...]
[Term]	comment: Generated 2nd	comment: Generated 1st
id: 000125847	[Term]	
name: ISO 10993-5	id: 000125884	[Term]
def: "ISO 10993-5 is a Cytotoxicity in-vitro screening test used to assess, in a fast and sensitive way, the biocompatibility of the test material when in contact ..." [URL ...]	name: Apoptosis Assay	id: 00042725
comment: Generated 2nd	def: "Apoptosis Assay is a detection and measurement system to monitor the occurrence of apoptosis in mammalian, anchorage-dependent cells ..." [URL ...]	name: Human Corneal Epithelium Model
[Term]	comment: Generated 1st	
id: 000125852	[Term]	[Term]
name: Agar Diffusion Method	id: 000125885	id: 00042726
def: "agar diffusion test, or the Kirby-Bauer disk-diffusion method is a means of measuring the effect of an antimicrobial agent against bacteria grown in culture." [URL ...]	name: Chemiluminescence Assay	name: EpiOcularTM
	def: "Chemiluminescence is the emission of light with limited emission of heat (luminescence), as the result of a chemical reaction. [In conclusion, the reference values established for the] chemiluminescence assay are applicable also for the enzyme-linked immunosorbent assay." [URL ...]	def: "EpiOcularTM is currently under validation as a Draize Replacement by ECVAM but since its introduction in 1985, EpiOcularTM has been used to determine the ocular irritancy of their products without using animals by many personal care and household product companies." [URL ...]
	comment: Generated 1st and 13th	comment: Generated 2nd
		[Term]
		id: 00042727
		name: SkinEthicTM HCE
		[Term]
		id: 00042729

10.2 Figures and Tables XXXI

name: SIFT	name: Silicon Microphysiometer Assay	[Term] id: 1576 name: Neurite Outgrowth Assay def: "Neurite Outgrowth Assay is a fully automated and validated software package for quantifying neurite outgrowth of neuronal cell cultures (cell lines and primary cells)." [URL ...] comment: Generated 1st
[Term] id: 00042730 name: Pig's Ear Test	[Term] id: 1397 name: Bovine Lens Culture Test	[Term] id: 1597 name: Mouse Ear Swelling Test def: "The mouse ear swelling test is a well-accepted method for quantitating the inflammatory response to contact irritants and sensitizing agents." [URL ...] comment: Generated 1st
[Term] id: 00042731 name: PREDISKIN TM	[Term] id: 1412 name: EYTEX def: "Eytex is an alternative testing method that evaluates eye irritancy of a protein alteration system by using an in vitro, or test tube, procedure." [URL ...] comment: Generated 1st	[Term] id: 1598 name: Mouse Ear Swelling Assay def: "noninvasive mouse ear swelling assay (MESA) for contact allergy testing was evaluated using fragrance components and complex fragrance mixtures." [URL ...] comment: Generated 3rd
[Term] id: 00042736 name: Cytosensor Microphysiometer	[Term] id: 1413 name: The Registry of Cytotoxicity def: "The Registry of Cytotoxicity : toxicity testing in cell cultures to predict acute toxicity (LD50) and to reduce testing in animals." [URL ...] comment: Generated 1st	[Term] id: 1608 name: Neutral Red Release Assay
[Term] id: 00042745 name: Reduced LLNA def: "Reduced local lymph node assay for skin allergy testing (reduction of 50 percent or more relative to conventional animal tests) Reduced Local Lymph Node Assay for skin allergy testing makes it possible to reduce animal use by up to 75 percent compared with traditional guinea pig and mouse tests." [URL ...] comment: Generated 1st and 2nd	[Term] id: 1426 name: TEA Assay	[Term] id: 1610 name: PREDISAFE TM
[Term] id: 00042756 name: BALB/c 3T3 NRU Assay def: "3T3 NRU are cytotoxicity assays in which the dye neutral red is taken up by living cells." [URL ...] comment: Generated 1st	[Term] id: 1427 name: Skin2 ZK1200	[Term] id: 1611 name: Neutral Red Uptake Cytotoxicity Assay def: "neutral red uptake cytotoxicity assay has been used to evaluate the effect of the photosensitizers on cell viability." [URL ...] comment: Generated 1st
[Term] id: 00042757 name: NHK NRU Assay	[Term] id: 1428 name: Skin2 ZK1000/1100	[Term] id: 1653 name: Immediate Early Gene Messenger RNA Measurement
[Term] id: 00042758 name: TER and PCP in Renal Cell Lines	[Term] id: 1429 name: MATREX	[Term] id: 1779 name: Comet Assay def: "comet assay is a versatile technique for detecting damage and with adjustments to the protocol can be used to quantify the presence of a wide variety of DNA altering lesions (damage). Comet assay is a valuable tool in genotoxicity testing as it detects a broad range of (primary) DNA damage in virtually any cell type, even in non-proliferating cells." [URL ...] comment: Generated 2nd and 5th
[Term] id: 00042846 name: Bacterial Gene Mutation Studies	[Term] id: 1431 name: Fluorescence Leakage Test	[Term] id: 1796 name: Neutral Red Uptake Inhibition Test
[Term] id: 00042848 name: In Vitro Cytogenicity Study	[Term] id: 1433 name: CAMVA Assay	[Term] id: 1806 name: In Vitro Skin Absorption Test def: "In Vitro Skin Absorption Test is a full replacement for the in vivo skin penetration test under OECD TG 428." [URL ...] comment: Generated 1st
[Term] id: 00042854 name: HPRT Test	[Term] id: 1467 name: Rodent Whole Embryo Culture Assay def: "whole embryo culture assay is endorsed as one of few good in vitro embryotoxicity assays available." [URL ...] comment: Generated 2nd	[Term] id: 1807 name: In Vitro Membrane Barrier Test Method
[Term] id: 00051067 name: Franz Cell Diffusion	[Term] id: 1477 name: Chicken Embryotoxicity Screening Test def: "Chicken Embryotoxicity Screening Test was used to estimate the beginning of the embryotoxicity dose range." [URL ...] comment: Generated 1st	[Term] id: 1808 name: Bovine Sperm Cell Assay
[Term] id: 00051069 name: Salmonella/Microsome Assay	[Term] id: 1534 name: Embryonic Stem Cell Test def: "embryonic stem cell test is an in vitro screening assay used to investigate the embryotoxic potential of chemicals by determining their ability to inhibit differentiation of embryonic stem cells into spontaneously contracting cardiomyocytes." [URL ...] comment: Generated 3rd	
[Term] id: 00051144 name: Aromatase Assay def: "Aromatase assay is a non-animal method that uses either human placental tissue or a human cell-line to detect substances that inhibit aromatase activity." [URL ...] comment: Generated 1st	[Term] id: 1564 name: Limb Bud Micromass Culture	
[Term] id: 102 name: Ames Test def: "Ames test is a useful test for carcinogenic substances which measures the ability of a substance to damage genetic material (DNA) in special strains of bacteria. Ames test is a biological assay used in genetics, generally genetic toxicology , to test for mutagenic properties of a chemical compound." [URL ...] comment: Generated 3rd and 7th	[Term] id: 1568 name: Bovine Sperm Cell Assay	
[Term] id: 1376		

XXXII 10 Appendix

name: Bacterial Reverse Mutation Test

[Term]
id: 1809
name: In Vitro Mammalian Chromosome Aberration Test
def: "In vitro mammalian chromosome aberration test is a short-term test used to identify structural chromosome aberrations in cultured mammalian cells." [URL ...]
comment: Generated 1st

[Term]
id: 1810
name: In Vitro Mammalian Cell Gene Mutation Test
def: "in vitro mammalian cell gene mutation test is a short term test for the evaluation of possible mutagenic effects of chemicals. in vitro mammalian cell gene mutation test can be used to detect gene mutations induced by chemical substances." [URL ...]
comment: Generated 1st and 2nd

[Term]
id: 1811
name: In Vitro Sister Chromatid Exchange Assay in Mammalian Cells

[Term]
id: 220
name: Local Lymph Node Assay
def: "local lymph node assay is an animal-based toxicology test developed as an alternative to the transdermal guinea pig sensitization test." [URL ...]
comment: Generated 1st

[Term]
id: 266
name: Bioluminescent Bacterial Genotoxicity Test

[Term]
id: 267
name: In Vitro Micronucleus Test
def: "in vitro micronucleus test is a valid and sensitive assay suitable for the detection of spindle poisons like paclitaxel and chromosome damaging toxicants like irradiation. In Vitro Micronucleus Test is a genotoxicity test system." [URL ...]
comment: Generated 2nd

[Term]
id: 268
name: In Vitro Cell Transformation Assay
def: "In vitro cell transformation assays are well established short-term predictive tests of tumorigenicity ." [URL ...]
comment: Generated 3rd

[Term]
id: 269
name: Cytotoxicity Testing
comment: Generated 1st

[Term]
id: 271
name: Frog Embryo Teratogenesis Assay - Xenopus
def: "- FETAX is a 4-day, whole embryo-larval developmental toxicity screening assay which uses young embryos of the South African clawed frog, *Xenopus laevis*. - FETAX is a 96-hour assay, which was developed to assess developmental toxicity, and has been used in both human health and ecological assessments. - FETAX is a simple test that rears recently fertilized *Xenopus laevis* embryos in a solution containing the potential teratogen. - FETAX is a validated method for developing models

alternative to the use of Mammals in research laboratories." [URL ...]
comment: Generated 1st

[Term]
id: 274
name: Mouse Lymphoma Assay
def: "mouse lymphoma assay is a genotoxicity test that provides information on a compounds potential to induce cancer events such as gene mutilations and chromosomal changes in vitro Mammalian Cell Micronucleus Test. The in vitro Mammalian Cell Micronucleus Test is a genotoxicity test that provides information on a compounds potential to induce chromosomal damage." [URL ...]
comment: Generated 1st

[Term]
id: 280
name: Yeast Mutagenicity Assay
def: "yeast mutagenicity assay are among those tests in common use for early toxicity screening." [URL ...]
comment: Generated 1st

[Term]
id: 295
name: Somatic Mutation and Recombination Assay

[Term]
id: 51
name: Fixed Dose Procedure
def: "The fixed-dose procedure is a more humane method to replace the LD50 in acute toxicity testing, which was first proposed by the British Toxicology Society (BTS, 1984)." [URL ...]
comment: Generated 1st

[Term]
id: 52
name: Up-And-Down Method
def: "up-and-down method is a procedure that has been confirmed to reduce the number of animals needed to determine LD50 values without compromising reliability." [URL ...]
comment: Generated 1st

[Term]
id: 53
name: Acute Toxic Class Method
def: "Acute-toxic-class method is an alternative to the classical LD50 test." [URL ...]
comment: Generated 4th

[Term]
id: 682
name: Acute Oral Toxicity Reduction Method

[Term]
id: 797
name: Transcutaneous Electrical Resistance Assay

[Term]
id: 805
name: CORROSITEX
def: "CORROSITEX is an in vitro test system that mimics the effect of corrosives on living skin while lowering testing costs and providing quicker results when compared to in vivo." [URL ...]
comment: Generated 2nd

[Term]
id: 806
name: Skin2 ZK1350

[Term]
id: 807
name: EPISKIN

def: "Episkin is a tri-dimensional human skin model with a reconstructed epidermis and a functional stratum corneum." [URL ...]
comment: Generated 7th

[Term]
id: 827
name: EpiDermTM
def: "EpiDermTM is a commercially available human skin model consisting of normal human-derived epidermal keratinocytes (NHEK), which have been cultured to form a multilayered, highly differentiated model of the human epidermis." [URL ...]
comment: Generated 1st

[Term]
id: 832
name: Red Blood Cell Photohemolysis / Hemoglobin Photooxidation Assay

[Term]
id: 837
name: Histidine Oxidation Assay

[Term]
id: 838
name: Candida Albicans Phototoxicity Test

[Term]
id: 839
name: Skin2 ZK1351

[Term]
id: 840
name: 3T3 NRU Phototoxicity Test
def: "3T3 NRU Phototoxicity Test is a modification of the BALB/c 3T3 NRU test and involves a shorter chemical exposure and the additional application of light. 3T3 NRU phototoxicity test makes such animal studies redundant, because it is an in vitro test: skin cells are exposed to ultraviolet light in a Petri dish. [The test results from the] 3T3 NRU phototoxicity test are used in a tiered testing approach to determine the phototoxic potential of test substances." [URL ...]
comment: Generated 1st

[Term]
id: 841
name: SOLATEX PI assay

[Term]
id: 842
name: NHK NRU PT Assay

[Term]
id: 843
name: EpiDerm Phototoxicity Assay

[Term]
id: 846
name: SKINTEX

[Term]
id: 851
name: Hydra Embryotoxicity Test

[Term]
id: 873
name: IRE Test

[Term]
id: 874
name: BCOP Test
def: "BCOP assay is the bovine corneal opacity and permeability assay, It uses isolated bovine corneas, a slaughterhouse by-product, instead of living rabbits." [URL ...]
comment: Generated 1st

PMID	Text passage
1453466	The structure ... has been solved by molecular replacement methods and <u>refined</u> at 1.45 Å
7811733	the AR enzyme will help in the <u>refinement</u> and design of future inhibitors.
8433965	Molecular <u>replacement</u> based on the wild type structure
8603861	of intraocular gene <u>replacement</u> therapy
8809072	multiple isomorphous <u>replacement</u> with two heavy-atom derivatives
9072292	massive <u>reductions</u> in total fat consumption with <u>replacement</u> with carbohydrates
9258368	the N-terminal amine <u>replacement</u> in combination with a 4-substituted phenacetyl
10089356	With poor isomorphous <u>replacement</u> experimental phases,
11558829	clinical trials to <u>refine</u> optimal management.
15299503	is applied to a molecular- <u>replacement</u> problem
15388930	molecular <u>replacement</u> , density modification and <u>refinement</u> from the output files
15572770	understanding maximum-likelihood <u>refinement</u> , molecular <u>replacement</u>
15929998	During the <u>refinement</u> process, we found acetate
16083332	Hormone replacement therapy
17084653	While enzyme <u>replacement</u> therapy ... substrate <u>reduction</u> therapy
17154614	Implication for replacement surfactant design from this work
17273852	<u>refine</u> the diagnosis-related group (DRG)
17387266	colorectal and stomach cancer and <u>reduce</u> mortality
18007059	molecular- <u>replacement</u> procedure and structural <u>refinement</u> is currently in progress.
18154369	significantly reduce the computational effort and systematically refine results
18243077	administer volume <u>replacement</u> to stabilize the patient.
18929066	we review potential metrics that might <u>refine</u> or <u>replace</u> present metrics
19148639	the reduction of operating costs by the <u>replacement</u> of an external carbon source
19191742	using normal <u>replacement</u> dosing

Table 10.11. Listing of 24 true negative documents which are not 3Rs relevant, see Section 8.5.3. All documents mention some synonym of term “3Rs Relevant”. The table provides one text passage per document overlapping with one synonym.

<p>[Term] id: 881 name: LVET def: "LVET is a modification of the Draize ocular irritation test that was developed by Griffith et al." [URL ...] comment: Generated 3rd</p> <p>[Term] id: 885 name: HET-CAM Test</p> <p>[Term] id: 886 name: CEET def: "Isolated Chicken Eye (ICE) Test Method for Identifying Ocular</p>	<p>Corrosives." [URL ...] comment: Generated 4th</p> <p>[Term] id: 888 name: RBC Test def: "RBC test is a biological in vitro test for rapid estimation of membrane and protein denaturing properties of surfactants." [URL ...] comment: Generated 2nd</p> <p>[Term] id: 896 name: Pollen Tube Growth Test def: "Pollen Tube Growth Test: In monitoring possible cytotoxic effects of</p>	<p>bioactive chemicals, it is desirable to have easy and sensitive test systems." [URL ...] comment: Generated 1st</p> <p>[Term] id: 897 name: Microtox def: "Microtox is a routine toxicity test which uses marine microorganisms (luminescence bacteria and algae cells). Microtox is a standardised toxicity test system which is rapid, sensitive, reproducible." [URL ...] comment: Generated 1st and 4th</p>
--	--	--

(a) 0.001					(b) 0.005				
Samples	Recall	Precision	F-measure		Samples	Recall	Precision	F-measure	
940	0.902	0.946	0.924		940	0.894	0.955	0.923	
938	0.887	0.954	0.919		938	0.859	0.957	0.906	
938	0.896	0.942	0.918		938	0.876	0.943	0.908	
938	0.910	0.949	0.929		938	0.902	0.953	0.927	
938	0.891	0.946	0.918		938	0.874	0.951	0.911	
(c) 0.01					(d) 0.001				
Samples	Recall	Precision	F-measure		Samples	Recall	Precision	F-measure	
940	0.862	0.964	0.910		940	0.764	0.978	0.858	
938	0.844	0.957	0.897		938	0.759	0.965	0.850	
938	0.864	0.946	0.903		938	0.768	0.955	0.851	
938	0.896	0.957	0.925		938	0.789	0.969	0.870	
938	0.859	0.955	0.905		938	0.776	0.960	0.858	
(e) 0.05					(f) 0.07				
Samples	Recall	Precision	F-measure		Samples	Recall	Precision	F-measure	
940	0.674	0.978	0.798		940	0.600	0.983	0.745	
938	0.665	0.978	0.792		938	0.586	0.982	0.734	
938	0.638	0.958	0.766		938	0.544	0.962	0.695	
938	0.674	0.966	0.794		938	0.584	0.968	0.729	
938	0.663	0.975	0.789		938	0.582	0.986	0.732	
(g) 0.1					(h) 0.15				
Samples	Recall	Precision	F-measure		Samples	Recall	Precision	F-measure	
940	0.526	0.996	0.688		940	0.443	0.995	0.613	
938	0.503	0.987	0.667		938	0.426	0.985	0.595	
938	0.480	0.978	0.644		938	0.412	0.975	0.579	
938	0.535	0.977	0.691		938	0.465	0.991	0.633	
938	0.499	0.987	0.663		938	0.429	0.985	0.597	
(i) 0.2									
Samples	Recall	Precision	F-measure						
940	0.340	1.000	0.508						
938	0.354	0.988	0.521						
938	0.339	0.981	0.504						
938	0.339	0.994	0.506						
938	0.386	0.989	0.555						

Table 10.12. Cross validations results for the classification of term “3Rs Relevant” with varying thresholds from 0.001 to 0.2.

Glossary

Many of the entries in this glossary have been generated or extended using the *DOG4DAG* definition extraction method which has been defined, developed, and evaluated as part of this thesis.


The generated glossary entries are displayed *italic* labeled with the symbol




3Rs principle

In the “The Principles of Humane Experimental Technique” Russell and Burch (1959) classified humane techniques under the headings of Replacement, Reduction, and Refinement – now commonly known as the three Rs.. 10, 207, 209

BETWEENESS centrality

 “Betweeness centrality is a measure devised to describe the fraction of shortest paths going through a given node, with high values indicating that a node can reach many other nodes.”. 128


Brill POS tagger

 “Brill POS tagger is a trainable rule-based part of speech tagger.”. XVI

CLOSENESS centrality

The closeness centrality is a measure devised to describe the mean distance of a node to all other directly connected nodes in the given graph.. 128



F-logic

 “F-logic is an object-oriented extension of predicate logic, which is particularly suitable for representing ontologies on the Semantic Web.”. 12

F-measure

For most evaluation tasks the performance of a system is usually a trade-off between precision and recall. Therefore, for an easier comparison of systems Van Rijsbergen (1979) introduced the F-measure, as the harmonic mean between precision and recall. The F-measure (F) is defined as the harmonic mean between precision and recall.. XV, XVIII, XX–XXIV, XXXV, 22, 23, 35, 36, 42, 44–47, 50–52, 55, 57, 58, 110, 119, 130, 141, 220, 223, 225, 237


Gene Ontology

 “Gene Ontology is a consortium project developed to create a list of biologically relevant and carefully structured terms that can be shared among all sorts of bioinformatics resources.”  “Gene Ontology is a collaborative effort to address the need for consistent descriptions of gene products in different databases.”. 12, 35, 75, 104, 128–130, 158, 185, 188, 198, 199


GermanNet

GermanNet is the German counterpart to WordNet.. XXIII


MEDLINE

 “Medline is a bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system and the preclinical sciences.”. XVII, 17, 43

MeSH

 “Medical Subject Headings is a huge controlled vocabulary (or metadata system) for the purpose of indexing journal articles and books in the life sciences.”. XVI, 17, 43, 128–130, 165, 188, 190, 193, 195, 198, 210, 211, 221


Natural Language Processing

 *"Natural Language Processing is an aspect of artificial intelligence in which the computer program translates human language input by matching input characters or sounds with information in the knowledge base."* XIX, XLI, 18, 19, 22, 23, 46


OBO format

The OBO format (representation language) is the text file format used by OBO-Edit, the open source, platform-independent application for viewing and editing ontologies. An OBO "[term]" entry requires as "id" and "name" to be specified. Optionally a definition can be specified as "def", also important synonyms as "synonym", distinguishing between "EXACT", "BROAD", "NARROW", or "RELATED". Four built-in relationship types exist, namely "is_a", "intersection_of", "union_of", "disjoint_from" for many entries, e.g. for definitions a "xrefs" can be specified as database reference (dbxref).. 12



OIL

 *"Ontology Inference Layer is an extension of RDFS."* 12


OWL

 *"OWL is a language for making ontological statements, developed as a follow-on from RDF and RDFS, as well as earlier ontology language projects including OIL, DAML and DAML+OIL."* 12, 167, 168


Open Biomedical Ontologies (OBO) Foundry

 *"Open Biomedical Ontologies (OBO) Foundry is an attempt to develop a suite of reference ontologies to support data integration across the entire domain of biomedical research."*  *"OBO ontologies are open, orthogonal, instantiated in the well-specified OBO syntax and designed to share a common name space."* 15

Part-Of-Speech

With *Part-Of-Speech* the grammatical classification, or the category a word is denoted, to which a word can be assigned to in the context of a phrase, sentence or paragraph. This categories can be many fold, but usually contain the classes noun, adjective, adverb, verbal, but also subject, object, and predicate.  *"part of speech is a particular grammatical class of word, for example noun, adjective, or verb."* XIII, XXI, XXIII, XXXVII, 27, 28, 30, 33, 37, 48, 51, 56, 61, 64, 66, 85, 86, 89–91, 101–103, 112, 114, 145, 146, 236


Pointwise Mutual Information

"The measure is a straightforward transformation of the independence assumption (on a specific point), $P(w_1, w_2) = P(w_1) * P(w_2)$, in a ratio. Positive values indicate that words occur together more than would be expected under the independence assumption. Negative values indicate that one word tends to appear only when the other does not. Values close to zero indicate independence." (by Terra and Clarke (2003))  *"Pointwise mutual information is a measure of association used in information theory and statistics....."* 38, 39



PubMed

PubMed is a service of the U.S. National Library of Medicine that includes over 20 million citations from MEDLINE and other life science journals for biomedical articles back to 1948. PubMed includes links to full text articles and other related resources (see <http://www.ncbi.nlm.nih.gov/pubmed/>).. XV, XVI, XXXVI, 17, 24, 75, 86, 91, 100, 101, 110, 128, 141, 148, 177, 188, 195, 198, 233


PubMed Central

 *"PubMedCentral is a repository of all the full-text links on PubMed."* XV, XVI

RDF

Resource Description Framework  *"RDF is a knowledge representation language dedicated to the annotation of resources within the framework of the semantic web."*  *"RDF is a data representation format for schema-free structured information that is gaining momentum in the context of Semantic-Web ..."* 12

RDFS

 *"RDFS is a set of primitives to describe lightweight ontologies in RDF (it uses the RDF model and syntax) and for RDF (the ontologies are used to type resources and relations)."* 12, 168

REACH

REACH is the new EU Chemicals Regulation (EC 1907/2006) and stands for **R**egistration, **E**valuation, **A**uthorisation, and **R**estriction of Chemicals. Once all regulations contained in REACH are in place, all companies manufacturing or importing chemical substances into the European Union in quantities of one tonne or more per year will be required to register these substances with a new European Chemicals Agency in Helsinki, Finland.. 207

Reduction

In the “The Principles of Humane Experimental Technique” Russell and Burch (1959) classified humane techniques under the headings of Replacement, Reduction, and Refinement – now commonly known as the three Rs.

Reduction means reducing the number of animals used to obtain information of a given amount and precision, or increasing the amount of useful data obtained from the same number of animals, without compromising the quality or the quantity of animal-based research. Three main ways for reducing animal use: a) better research strategy; b) better control of variation; c) better statistical analysis.. 10

Refinement

In the “The Principles of Humane Experimental Technique” Russell and Burch (1959) classified humane techniques under the headings of Replacement, Reduction, and Refinement – now commonly known as the three Rs.

Refinement means any decrease in the severity of inhumane procedures applied to those animals which still have to be used.. 10

Replacement

In the “The Principles of Humane Experimental Technique” Russell and Burch (1959) classified humane techniques under the headings of Replacement, Reduction, and Refinement – now commonly known as the three Rs.

Replacement means that higher order animals, which are capable of suffering or feeling pain, should not be used if the aims in research, teaching, or testing can be achieved in other ways. 10


TREC

Text REtrieval Conference (TREC) is a conference which had a question answering track since 1999; in each track the task was defined such that the systems were to retrieve small snippets of text that contained an answer for open-domain, closed-class questions (i.e., fact-based, short-answer questions that can be drawn from any domain). See <http://trec.nist.gov/data/qa.html>.. XVIII, 50



TextTree

The TextTree is a representation for text in a tree structure. Each node represents a non-overlapping sub-string of the text. Nodes correspond to tagged regions, e.g. tokens, sentence. Nodes can hold several types of tags, e.g. tokens will have Part-Of-Speech tags if they represent a word.. 63


UMLS

The Universal Medical Language System (UMLS) is a large multi-lingual vocabulary database, that contains biomedical and health related concepts and their various names. The vocabulary is categorized by semantic types e.g. Chemicals & Drugs, Anatomy (description from Tsuruoka et al. (2008)).  “UMLS is a system designed by the National Library of Medicine (NLM) to help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of bibliographic databases, factual databases, and expert systems.”. XV, XVI, XIX, 17, 33, 40, 47, 63, 142

WordNet

 “Wordnet is a databse of the English language that is lexical and it assists search engines to understand the relationship between words.”  “WordNet is a powerful lexical reference system that combines aspects of dictionaries and thesauri with current psycholinguistic theories of human lexical memory.”. XXII, 36, 37, 40, 50

WordNet glosses

 “Wordnet is a databse of the English language that is lexical and it assists search engines to understand the relationship between words.” WordNet glosses, the entries in WordNet, are composed of two parts, a definition part and a sample part.. XVIII, 46

ZEBET

ZEBET is the Centre for Documentation and Evaluation of Alternative Methods to Animal Experiments in the Federal Institute for Risk Assessment (BfR) Berlin, Germany. 177


abbreviation

An *abbreviation* is the short form of a word or word phrase.. XVII, 32, 40, 63


accuracy

Accuracy is defined as the percentage of items selected right (true positives + true negatives) and the corresponding error is defined as the percentage of wrongly selected items (false positives + false negatives).. XVII, 23, 43, 50, 55


acronym

 “acronym is a word formed from the initial letters or syllables taken from a word or group of words that forms a shorter, abbreviated term representing the original idea.”. XVII, 32, 33, 41–43, 56, 63


adjective

 “adjective is a word which qualifies a noun, that is, shows or points out some distinguishing mark or feature of the noun.”. XIII, XXXVI, 19, 40, 66

adverb

 “adverb is a word or clause that typically describes or modifies a verb (He ate noisily), but can also modify an adjective (She is extremely short) or another adverb (He sang exceptionally poorly).”. XXXVI, 19, 40, 66


anaphora

 “Anaphora is a rhetorical device that consists of repeating a sequence of words at the beginnings of neighboring clauses, thereby lending them emphasis.”. XVIII, 46

automatic term recognition

Automatic term recognition is the extraction of domain relevant terminology from natural language text using linguistic and statistical information. Regarding ontology learning, ATR is the first of the ontology learning sub tasks.. 62


average precision

 “Average precision is a common evaluation measure for system rankings, and is computed as the average of the system’s precision values at all points in the ranked list in which recall increases, that is at all points in the ranked list for which the gold standard annotation is YES (Voorhees and Harman, 1999).”. 23, 45, 76, 82, 83


class

Classes provide an abstraction mechanism for grouping resources with similar characteristics. In the scope of this work class is used synonymous with concept.. 11, 14, 37, 121

collocation

Within the area of corpus linguistics, collocation is defined as a sequence of words or terms which co-occur more often than would be expected by chance (from <http://en.wikipedia.org/wiki/Collocation>  “Collocation is an aspect of language generally considered arbitrary by nature and problematic to L2 learners who need collocational competence for effective communication.”). XIV, 32

combinatorics

In statistics, *combinatorics* is a branch of pure mathematics concerning the study of discrete (and usually finite) objects. It is related to many other areas of mathematics, such as algebra, probability theory, ergodic theory and geometry, as well as to applied subjects in computer science and statistical physics. (Wikipedia)  “Combinatorics is an area of discrete mathematics that studies collections of distinct objects and the ways that they can be counted or ordered, or used to satisfy some optimality criterion.”. XL, 96

concept

The *concept*, as used here, groups a number of terms, corresponding synonyms, and abbreviations to a semantic unit, which can be referred to by all assigned terms. Concepts are defined by a natural language definition, and have a representative label (usually but not necessarily identical with one of the terms). In the scope of this work concept is used synonymous with *class*. XXIII, XXIV, XXXVIII, XLII, 11, 26, 33, 37, 54, 55, 63, 66, 69, 121, 160, 169, 189, 205


conditional probability

The *conditional probability* describes the probability of an event A under the condition of the occurrence of some other event B , written $P(A|B)$.. XIX, 47, 50, 79


consistency

Like databases, an ontology is consistent if no integrity constraint is violated in its current state. The language specification, e.g. for OWL (Patel-Schneider and Horrocks, 2004), is a source for consistency constraints.. 169

controlled vocabulary

A controlled vocabulary is a structured set of terms and definitions agreed by an authority or community (Spasić et al., 2008).  “controlled vocabulary is an established list of words and phrases (generally referred to as subject headings or descriptors) that provides a standard vocabulary used in a database to describe the various items in that database.”. XV, 11

cosine similarity

 “Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining.”. XXIII, 51, 128


definiendum

In definition extraction or definitional question answering, the *definiendum* is the unknown term to be defined (Storrer and Wellinghoff, 2006).. XVIII, XIX, 45, 108, 110–112, 163


definiens

In definition extraction or definitional answering question, the *definiens* describes the meaning postulated for the term (Storrer and Wellinghoff, 2006).. XIX, 108, 110, 112, 163

definitional question

A definitional question is a question containing only the term to be defined, e.g. “What is a rat?” or “Who is Thomas Wächter?”.  “definitional question is a question of the type such as but not limited to “What is X?”, “Who is Y?”, etc.”. XVIII, XIX, XXXVIII, XXXIX, 44–46


definitional question answering

Definitional question answering is a subtask of question answering where definitional questions need to be answered. A definitional question is a question containing only the term to be defined, e.g. “What is a rat?” or “Who is Thomas Wächter?”.  “Definitional question answering is a task of answering definitional questions used for finding out conceptual facts or essential events about the question target.”. XXXIX, 43, 57

definitior

In definition extraction or definitional question answering, the *definitior* denotes the verb, which relates the definiens component to the definiendum component (Storrer and Wellinghoff, 2006).. XIX, 47, 112

differentia

In the context of ontologies the *differentia* describes the set of properties that distinguish the term from other members of the class (definition by the Gene Ontology Consortium).  “differentia is a difference between two things.”. XXI, 11, 14, 111


discrete probability distribution

The discrete probability distribution arise in the mathematical description of probabilistic and statistical problems in which the values that might be observed are restricted to being within a pre-defined list of possible values. This list has either a finite number of members, or at most is countable.. XL, 96


document frequency

The *document frequency* denotes “the number of documents that make use of the term. When a term occurs in more documents, then it is less important for the purpose of information retrieval.” (Baclawski and Niu, 2005) Frequent terms are much less selective than rare terms.. XLIII

entropy

In information theory, entropy is a measure of the uncertainty associated with a random variable. (from Wikipedia)  “Entropy is an information theoretical concept applied across physics, information theory, mathematics and other branches of science and engineering.”. XL, 22, 36

factoid question

A factoid question is a fact-based, short answer question such as “How many people work at TU Dresden?”.  “factoid question is a fact-based, short answer question such as “How many calories are there in a Big Mac?”. XVIII, XXXIX, 44

false negative

From the overlap of the items selected by some system, the false negatives are the expected items the method missed to select.. XVI, XXXIX, 22, 25

false positive

From the overlap of the items selected by some system, the false positives are the wrongly selected items.. XVI, XXXIX, 22, 25

finite state grammar

“A finite state grammar models a sentence as a succession of ‘states’ progressing left-to-right. Each word chosen determines what can follow it. Chomsky proved that such a grammar could generate an infinite number of sentences, but could not cope with other aspects of language such as discontinuous structure.”, excerpt from Chomsky (1957). XIX, XXXIX

genus

The genus denotes the category or kind a concept belongs to. E.g. Stem cell is a *kind of cell*.. XXI, 11

global abbreviation

Global abbreviations are such abbreviations which occur in documents *without* their long forms.. 40

gold standard

A *gold standard* can be a data set produced by a method that is widely accepted as being the best available or it can be manually created by experts to compare different systems.. 24

holonym

Holonymy (in Greek holon = whole and onoma = name) is a semantic relation. Holonymy defines the relationship between a term denoting the whole and a term denoting a part of, or a member of, the whole.(<http://en.wikipedia.org/wiki/Holonymy>); Example: “cell membrane” is a holonym to “cell”, because the “cell membrane” is part of the “cell”. ¹⁹₀₀₀₁₀ “*holonym is a meronym’s opposite, namely a word or combination of words that designate a thing or concept which includes some other thing or concept, as per ATOM, a holonym in relation to PROTON.*”. XIX

hypergeometric distribution

The Hypergeometric distribution is a discrete probability distribution used in combinatorics. Assuming a finite number of elements, randomly n elements get drawn without replacement. The hypergeometric distribution describes the probability of drawing an element with the requested property.. XL, 66, 96, 100

hypernym

“A hypernym is a word with a general meaning that has basically the same meaning of a more specific word.” (Laurie Beth Feldman, Morphological Aspects of Language Processing, Lawrence Erlbaum Associates, 1995); Example: “cell” is hypernym to “stem cell”, because a “stem cell” is a special type of “cell”. ¹⁹₀₀₀₁₀ “*hypernym is a linguistic term for a word whose meaning includes the meanings of other words, as the meaning of transportation includes the meaning of train, chariot, dogsled, airplane, and automobile.*”. XV, XIX, XXI–XXIII, XLII, 34, 37, 38, 50, 51, 58

hyponym

In linguistics, a hyponym is a word or phrase whose semantic range is included within that of another word. <http://en.wikipedia.org/wiki/Hyponym>; Example: “liver” is hyponym to “organ”, because “liver” is a “organ”. XIX, XXI–XXIII, XL, XLII, 37, 38, 51, 58, 108, 112

instance

Instances are the objects which have been categorised to belong to one class. According to OWL Web Ontology Language Reference, the individuals in the class extension are called the *instances* of the class.. 11, 14, 169

learning accuracy

The measure was introduced by Hahn and Schnattinger (1998). It “measures not only the overall correctness of the final classification but also incorporates the distance between the position *f* predicted by the algorithm and the correct one *s*.” (see Witschel (2005)). XXIII, 23

list question

A list question is a question for which there exist multiple correct answers, e.g. “In which country is Dresden located?”. XVIII, XL, 44

local abbreviation

Local abbreviations are such abbreviations which occur in documents together *with* their long forms.. 40

machine learning

¹⁹₀₀₀₁₀ “*Machine learning is an area of computer science concerned with the development of data-driven approaches and algorithms addressing the fundamental problem of finding and describing complex structure in high-dimensional, non-stationary, non-linear and noisy data.*”. XVII, XXI, XXII, 38, 40, 43, 47, 48, 50, 58, 59, 110


maximum entropy

The *maximum entropy* method was introduced by Berger et al. (1996) and is a method for statistical modelling where minimal assumptions are made about the data. The method allows the assignment of a-priori probability to known classes based on incomplete information. As the name suggest, the method aims to maximize the entropy and the authors describe the methods goal as follows: *model all that is known and assume nothing about that which is unknown. In other words, given a collection of facts, choose a model consistent with all the facts, but otherwise as uniform as possible.*. XVIII, 21, 22, 46






n-gram

¹⁹₀₀₀₁₀ “*ngram is an ordered sequence of n adjacent words, characters, or morphological adornments.*”. XVI, 43



named entity recognition

 "Named Entity Recognition is an information extraction task that is concerned with finding textual mentions of entities that belong to a predefined set of categories.". 35

noun

 "noun is a kind of word (see part of speech) that is usually the name of a person, place, thing, quality, or idea."  "noun is a word used to refer to people, animals, objects, substances, states, events and feelings."  "Noun is a word that names beings, things, places, qualities, actions or states."  "noun is a word used to name a person, animal, place, thing, and abstract idea."  "noun is a word that names a person, animal, place, thing, idea, or concept.". XIII, XIX, XXII, XXIII, XXXVI, 19, 40, 51, 66


noun phrase

 "noun phrase is a word or group of words, which acts as the subject, complement or object of a clause, or as the object of a preposition."  "A noun phrase is a collection of words that functions together as a unit, with the noun identifying the core actor or recipient of the action.". XIX, 19, 20, 51, 55, 63, 64, 66, 111, 163



object

An object in grammar is a sentence element and is often part of the sentence predicate. It denotes somebody or something involved in the subject's "performance" of the verb. (Wikipedia). XXXVI

object complement

An object complement is an noun, pronoun, or adjective which follows a direct object and re-names it or tells what the direct object has become. It is most often used with verbs of creating or nominating such as make, name, elect, paint, call, etc. (from <http://englishplus.com/>  "object complement is a word (often an adjective, noun or pronoun) or a phrase which follows an object in a sentence to add more information to it or to describe its recent condition.". XXIV


ontology

 "Ontologies are a widely accepted state-of-the-art knowledge representation, and have thus been identified as the central enabling technology for the Semantic Web."  "Ontology is a new concept that is emerging from the various semantic Web initiatives, which roughly speaking can be defined as a semantic system that contains terms, the definitions of those terms, and the specification of relationships among those terms.". 11, 189


ontology learning

"Ontology Learning" was introduced by Maedche and Staab (2001) and described as the acquisition of a domain model from data.. XLII, 8, 19, 26, 35, 55, 108, 122


participle phrase

 "participle phrase is a phrase containing a participle and any complements or modifiers it may have.". XIX, 20

phrase chunking

 "Phrase chunking is a natural language process that separates and segments sentences into their subconstituents, such as noun, verb, and prepositional phrases.". 19

precision

The measure precision is defined as the proportion of selected items that a method selected correctly.  "Precision is a measure of tests reproducibility when repeated on the same sample.". XIII, XIV, XVI-XIX, XXI, XXIII, XXIV, 22, 73, 74, 79, 81, 90, 102, 220


predicate

In traditional grammar, a predicate is one of the two main parts of a sentence (the other being the subject, which the predicate modifies). (Wikipedia). XXXVI


preposition phrase complement

A preposition phrase complement is a required by some verbs like borrow or depend. Example: The lender borrows a book. A ski resort depends on snow. where borrows_{subj} = lender and borrows_{obj} = book or depends_{subj} = resort and depends_{obj} = snow.. XXIV

pronoun


 "pronoun is a function word that is used in place of a noun or noun phrase.". 40

pronoun phrase


 "pronoun phrase is a phrase whose head is a pronoun.". XIX, 20

property restriction


The web ontology language (OWL) allows to restrict the individuals of a class by using restrictions of some property. E.g. if a cell cycle process like mitosis only occurs in cell having a nucleus a restriction can be formulated that only individuals which are eucaryotes can participate in mitosis.

 "Property Restriction is a way of defining a Class by restriction the behaviour of a Property.". 108

recall

The measure recall is defined as the proportion of items a method should have selected.  *"Recall is a measure of completeness."*. XIV, XVI–XIX, XXI, XXIII, XXIV, 23, 73, 74, 220

sentence splitting

Sentence splitting is the Natural Language Processing technique used to separate a document containing natural language text into the sequence of sentences.  *"Sentence splitting is a necessary pre-processing step for a number of Natural Language Processing (NLP) tasks including part-of-speech tagging and parsing."*. XVIII, XIX, 19, 46


similarity

Similarity is a measure which quantifies the common and distinct properties of some objects.. XXIV, 52


snippet

A *snippet* denotes a short textual summary typically returned as search result from keyword based search engines. The snippet contains sections from the whole document which contain the key-words.. XVIII, 46, 111, 161

specificity

 *"Specificity is an important measure of the quality of a test and an indication of how well the test performs in excluding disease (see ROC curve, sensitivity)."*. XXIV, 52


stemming

 *"Stemming is an algorithm that determines how to find the root form of words based on the linguistic characteristics of the language."*. 19


subject

The subject has the grammatical function in a sentence of relating its constituent (a noun phrase) by means of the verb to any other elements present in the sentence, i.e. objects, complements and adverbials. (Wikipedia). XXXVI


subject complement

A subject complement is a phrase following the linking verb that renames or describes the subject of a sentence. Example: *I was a PhD student*. "PhD student" is linked through the verb was to the subject "I".  *"subject complement is a word or group of words that completes the meaning of a linking (intransitive) verb and modifies (or refers to) the subject of a clause."*. XXIV


superclass

A *superclass* describes a class from which other classes are derived. In ontology engineering a *superclass* can be seen as the parent concept in a relation. In the specific case of hyponym relations, the hyponym is the *superclass*.  *"superclass is an object one-level higher in the hierarchy than an object and a subclass is an object one-level below."*. 11


supervised learning

 *"Supervised learning is a kind of machine learning where the learning algorithm is provided with a set of inputs for the algorithm along with the corresponding correct outputs, and learning involves the algorithm comparing its current actual output with the correct or target outputs, so that it knows what its error is, and modify things accordingly."*. 39


support vector machine

 *"Support Vector Machine is a statistical analysis technique which generates a separator to divide the document space into several regions, each corresponding to a specific author."*. XVII, 43


synonym

 *"synonym is a word (or sometimes phrase) which has the same meaning (or almost the same meaning) as another word in the same language."*. 32


syntactic pattern

In linguistics, syntactic patterns are a form of model using rules regarding the grammatical structure of a sentence.  *"syntactic pattern is a global feature in that it is found in practically all varieties of English world wide, but it is used with different frequencies in different Englishes and also shows differences at the micro-variational level (co-occurrence patterns, choice of subject, etc.)"*. XIX

t-test

 *"t-test is an inferential test that measures whether random sampling alone is the reason for group differences (Nelson, 1981)."*. XIV, 32

taxonomy

 *"Taxonomy is the science of classification according to a pre-determined system, whose resulting catalogue is used to provide a conceptual framework for discussion or analysis."*. 12

taxonomy generation

In ontology learning, *taxonomy generation* is the process of the automatic prediction of subsumption relationship, hypernym-hyponym relationships.. 122


term

By terms we refer to phrases from natural language which can be simple nouns like “cell” or “growth”, or noun phrases like “early endosome”, “epidermal growth factor” which are essentially single grammatical units containing a noun as main word, here “endosome” and “factor”. More complex terms can be composed from several noun phrases like “endosomal sorting complex required for transport proteins” or “transcription factors involved in the regulation of endocytosis”. XLII, 11, 27, 32, 33


term frequency

The *term frequency* denotes “the number of times that the term occurs in a document. The assumption is that if a term occurs more frequently in the document, then it must be more important.” (Baclawski and Niu, 2005). XLIII


tf-idf

term frequency inverse document frequency (tf-idf) is an often used statistical measure in information retrieval.  “TFIDF is an algorithm for extracting feature words (words recognized as important) in sentences of a document, and is used in fields such as an information retrieval.”. XVI, XVIII, 28, 45, 46, 64, 67, 75, 79, 81, 84, 99, 100, 102, 146, 236


thesaurus

 “thesaurus is a book that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which contains definitions and pronunciations.”. 12



token

 “Token is a basic graphemic unit found in a text.”. XLIII, 19

tokenization

tokenization is a processing steps where a text gets separated in word or other meaningful smallest syntactical units of text (tokens).  “Tokenization is a fundamental pre-processing step in Information Retrieval systems in which text is turned into index terms.”. XIX, 46

topic map

 “Topic Map is a new tool proposed by the International Organization for Standardization (ISO) to solve problems about knowledge representation and knowledge management.”  “topic map is a collection of information organized around binding points (topics), which makes topic map information amenable to interchange in fragments.”. 12

true negative

From the overlap of the items selected by some system, the true negatives are the items not selected which were not expected to be selected.. XIV, XLIII, 22, 32


true positive

From the overlap of the items selected by some system, the true positives are the correctly selected items which can be expected to become selected.. XIV, XLIII, 22, 25


validity

Regarding validity it can be distinguished between syntactically and semantically valid ontologies. Syntactically correct ontologies declare all concepts and relations used whereas semantically valid ontologies use all concepts or relations as declared for the ontology.. 169


verbal

 “Verb is a part of speech consisting of a word or group of words that signify an action, condition or experience.”. XIII, XIX, XXXVI, 19, 40, 47, 66

verbal phrase

 “verbal phrase is a composition of a postpositive phrase and a verb, whereas the postpositive phrase is being used to limit or otherwise modify the verb at hand.”. XIX, 20

word sense disambiguation

 “Word Sense Disambiguation is a subtask of semantic tagging, which consists of assigning a semantic class (sense) to a lexical item as specified by a semantic lexicon.”. 21, 35, 36