2012

Knowledge Acquisition in a System

Christopher J. Thomas
Wright State University

Follow this and additional works at: http://corescholar.libraries.wright.edu/etd_all
Part of the Computer Engineering Commons, and the Computer Sciences Commons

Repository Citation

Thomas, Christopher J., "Knowledge Acquisition in a System" (2012). Browse all Theses and Dissertations. Paper 651.

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
Knowledge Acquisition in a System

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

by

Christopher J. Thomas
B.S., Universität Koblenz

2012
Department of Computer Science and Engineering
Wright State University
I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Christopher J. Thomas ENTITLED Knowledge Acquisition in a System BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy in Computer Science.

Amit P. Sheth, Ph.D.
Thesis Director

Mateen Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Andrew Hsu, Ph.D.
Dean, Graduate School

Amit P. Sheth, Ph.D.
Pankaj Mehra, Ph.D.
Shaojun Wang, Ph.D.
Pascal Hitzler, Ph.D.
Gerhard Weikum, Ph.D.
ABSTRACT

Thomas, Christopher. PhD, Department of Computer Science and Engineering, Wright State University, 2012. Knowledge Acquisition in a System.

I present a method for growing the amount of knowledge available on the Web using a hermeneutic method that involves background knowledge, Information Extraction techniques and validation through discourse and use of the extracted information.

I present the metaphor of the “Circle of Knowledge on the Web”. In this context, knowledge acquisition on the web is seen as analogous to the way scientific disciplines gradually increase the knowledge available in their field. Here, formal models of interest domains are created automatically or manually and then validated by implicit and explicit validation methods before the statements in the created models can be added to larger knowledge repositories, such as the Linked open Data cloud. This knowledge is then available for the next iteration of the knowledge acquisition cycle.

I will both give a theoretical underpinning as well as practical methods for the acquisition of knowledge in collaborative systems. I will cover both the Knowledge Engineering angle as well as the Information Extraction angle of this problem. Unlike traditional approaches, however, this dissertation will show how Information Extraction can be incorporated into a mostly Knowledge Engineering based approach as well as how an Information Extraction-based approach can make use of engineered concept repositories. Validation is seen as an integral part of this systemic approach to knowledge acquisition.

The centerpiece of the dissertation is a domain model extraction framework that implements the idea of the “Circle of Knowledge” to automatically create semantic models for domains of interest. It splits the involved Information Extraction tasks into that of Domain Definition, in which pertinent concepts are identified and categorized, and that of Domain Description, in which facts are extracted from free text that describe the extracted concepts. I then outline a social computing strategy for information validation in order to
create knowledge from the extracted models.

This dissertation makes the following contributions:

• A hermeneutic methodology for knowledge acquisition within a system, involving
  – Human and artificial agents
  – Formally represented knowledge,
  – Textual information,
  – Information Extraction methods and
  – Information validation techniques

• Ontology Design

• Automatic Domain Model creation
  – Top-down Domain hierarchy extraction (Domain Definition)
  – Bottom-up Pattern-based extraction of named relationships (Domain Description)
    * Distantly supervised Relational Targeting Information Extraction
    * Probabilistic positive-only Multi-class classifier
    * Statistical measure for relationship pertinence
    * Recall enhancement using pattern generalization

• Implicit and Explicit Information validation
## Contents

1 Introduction ................................. 1
  1.1 Motivation ...................................... 2
  1.2 Hypothesis ..................................... 4
  1.3 Scope .......................................... 11

2 Overview ..................................... 13
  2.1 Terminology ................................... 16
  2.2 Knowledge Engineering oriented knowledge acquisition ...... 18
  2.3 Information Extraction oriented knowledge acquisition ..... 20
    2.3.1 Epistemological Considerations ................... 21
    2.3.2 Automatic Domain Model Creation ............... 23

3 Epistemological Foundations .............. 29
  3.1 Introduction .................................. 29
  3.2 Knowledge .................................... 30
    3.2.1 Truth ..................................... 34
    3.2.2 Justification ................................. 40
    3.2.3 Belief ..................................... 41
    3.2.4 Knowledge in a Group - Social Epistemology .... 41
    3.2.5 Knowledge in a System - Systems Epistemology ... 43
  3.3 Reference .................................... 47
    3.3.1 Rigid Designators ............................ 48
    3.3.2 Definite Descriptions ....................... 48
    3.3.3 Application ................................ 49
  3.4 The Hermeneutic Circle ...................... 51
  3.5 Knowledge Acquisition in a system ...................... 52
    3.5.1 Practical Considerations .................... 55

4 Knowledge Engineering - Based Domain Model Creation ....... 57
  4.1 Introduction .................................. 58
  4.2 Ontology Design .............................. 60
    4.2.1 General Considerations ..................... 60
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.2 Domain Definition - Schema Design</td>
<td>63</td>
</tr>
<tr>
<td>4.2.3 Archetypal Instances</td>
<td>65</td>
</tr>
<tr>
<td>4.2.4 Application example</td>
<td>71</td>
</tr>
<tr>
<td>4.2.5 Instances as Archetypes of Concepts</td>
<td>72</td>
</tr>
<tr>
<td>4.2.6 Implications</td>
<td>75</td>
</tr>
<tr>
<td>4.3 Populating the Ontology</td>
<td>75</td>
</tr>
<tr>
<td>4.3.1 General Considerations</td>
<td>75</td>
</tr>
<tr>
<td>4.3.2 Populating GlycO from trusted sources</td>
<td>76</td>
</tr>
<tr>
<td>4.3.3 An Intelligent Population Algorithm</td>
<td>77</td>
</tr>
<tr>
<td>4.4 Evaluation</td>
<td>79</td>
</tr>
<tr>
<td>4.5 Conclusion</td>
<td>81</td>
</tr>
<tr>
<td>5 Automatic Domain Model Extraction</td>
<td>83</td>
</tr>
<tr>
<td>5.1 Background</td>
<td>88</td>
</tr>
<tr>
<td>5.1.1 Domain Definition</td>
<td>88</td>
</tr>
<tr>
<td>5.1.2 Domain Description</td>
<td>90</td>
</tr>
<tr>
<td>5.2 Related Work</td>
<td>95</td>
</tr>
<tr>
<td>5.2.1 Ontology Learning</td>
<td>95</td>
</tr>
<tr>
<td>5.2.2 Top-Down Extraction of Knowledge</td>
<td>96</td>
</tr>
<tr>
<td>5.2.3 Bottom-up Extraction of Knowledge</td>
<td>96</td>
</tr>
<tr>
<td>5.3 Domain Definition - Hierarchy Creation</td>
<td>102</td>
</tr>
<tr>
<td>5.3.1 Expansion</td>
<td>104</td>
</tr>
<tr>
<td>5.3.2 Reduction</td>
<td>108</td>
</tr>
<tr>
<td>5.3.3 Synonym Acquisition</td>
<td>110</td>
</tr>
<tr>
<td>5.3.4 Serialization</td>
<td>111</td>
</tr>
<tr>
<td>5.4 Domain Description - Information Extraction</td>
<td>112</td>
</tr>
<tr>
<td>5.4.1 Surface Patterns</td>
<td>114</td>
</tr>
<tr>
<td>5.4.2 Probabilistic Framework</td>
<td>116</td>
</tr>
<tr>
<td>5.4.3 Vector-Space model</td>
<td>120</td>
</tr>
<tr>
<td>5.4.4 Pertinence</td>
<td>123</td>
</tr>
<tr>
<td>5.4.5 Relationship Domain and Range probabilities</td>
<td>128</td>
</tr>
<tr>
<td>5.4.6 Matrix-Based Fact Extraction</td>
<td>130</td>
</tr>
<tr>
<td>5.4.7 Pattern Analysis</td>
<td>131</td>
</tr>
<tr>
<td>5.4.8 Discussion</td>
<td>134</td>
</tr>
<tr>
<td>5.5 Model Completion - Combining Definition and Description</td>
<td>134</td>
</tr>
<tr>
<td>5.5.1 Concept Pairing heuristic - Wikipedia</td>
<td>135</td>
</tr>
<tr>
<td>5.5.2 Concept Pairing heuristic - General Case</td>
<td>135</td>
</tr>
<tr>
<td>5.5.3 Model-Creation</td>
<td>136</td>
</tr>
<tr>
<td>5.6 Evaluation</td>
<td>137</td>
</tr>
<tr>
<td>5.6.1 Hierarchy Creation Evaluation</td>
<td>137</td>
</tr>
<tr>
<td>5.6.2 Fact Extraction Evaluation</td>
<td>147</td>
</tr>
<tr>
<td>5.6.3 Evaluation of the full Domain Models</td>
<td>157</td>
</tr>
<tr>
<td>5.6.4 Discussion</td>
<td>163</td>
</tr>
</tbody>
</table>
# 6 Knowledge Verification and Propagation

6.1 Introduction .................................................. 167
   6.1.1 Explicit Validation ........................................ 169
   6.1.2 Validation in Use ........................................ 169

6.2 Discussion ................................................... 172
   6.2.1 User feedback to focused browsing ...................... 172
   6.2.2 Qualitative Evaluation of browsed facts ............... 172

6.3 Propagation of validated statements .......................... 174

6.4 Conclusion .................................................... 175

# 7 Conclusion .................................................... 177

7.1 Outlook ........................................................ 181

Bibliography ....................................................... 182

A Appendix A ....................................................... 211
List of Figures

1.1 Circle of (Web knowledge) life .................................. 5
2.1 Classification of the work in this dissertation in terms of Knowledge Engineering vs. Information Extraction .................................................. 15
2.2 Traditional Ontology Learning Layer Cake .................................. 25
2.3 Doozer++ Ontology Learning Layer Cake .................................. 25
3.1 Circle of (Web knowledge) life .................................. 53
3.2 Nonaka’s Knowledge Spiral .................................. 55
4.1 Selection of the first 3 levels in the GlycO hierarchy ......................... 66
4.2 Part of the relationship hierarchy .................................. 67
4.3 The GlycoTree structure that subsumes most known N-Glycans. ............ 69
4.4 Snapshot of the GlycO Pathway Browser .................................. 72
4.5 GlycO population workflow. .................................. 78
4.6 A part of the N-Glycan biosynthesis pathway as encoded in GlycO. For better visibility, only few relationship types are visualized. N-glycan_b-D-GlcNAc_13 is the beta-D-GlcNAc residue number 13 as enumerated in the GlycoTree model. .................................. 81
5.1 Small excerpt of a connected concept graph. For better visualization, the class hierarchy has been flattened. .................................. 88
5.2 Positioning of the Information Extraction portion of this work. Doozer++ is positioned in the no-NLP, many types of relationships corner. .................. 99
5.3 Steps 1 (Full-text search) and (Semantic Similarity) in the expansion process .................................. 105
5.4 Link types in WikiPedia .................................. 107
5.5 Network representation of the classifier. .................................. 119
5.6 Comparing precision and recall of fact extraction with and without pertinence. Pertinence has most influence in high-recall regions. Intuitively, as the confidence threshold is increased, patterns that are highly indicative of specific relationships contribute more to the classification and thus the impact of pertinence is diminished. .................................. 128
5.7 $F_1$ measures, computed against a reduced glossary, for the lists of terms generated by various mining tools .................................. 143
5.8 The Oncology subtree of the created MeSH-related category hierarchy. The size of the rectangles indicates the number of descendants. Lighter rectangles indicate categories, darker rectangles indicate individuals.

5.9 Evaluation wrt. the Wikipedia subset of terms in MeSH versions of 2004 and 2008.

5.10 Precision and recall: DBPedia test set and Wikipedia text corpus.

5.11 Precision and recall: UMLS test set and MedLine text corpus.

5.12 Precision on the UMLS and DBPedia testing sets averaging over all extracted facts.

5.13 Support vs. precision of relationships. The left scale indicates the training examples, the right one indicates precision for a relationship type.

5.14 Precision/Recall curve for select relationships, comparison between Doozer++ and the POS-version as well as the parse-version of WOE.

5.15 Comparison between Doozer++ and Mintz over all DBPedia relationships. (P) indicates a precision-oriented experiment without pattern generalization, (R) indicates a recall-oriented experiment with a generalization factor of 2.

5.16 Precision of the fact extraction for the on-demand created Models, depending on the confidence threshold $\epsilon_{rel}$.

5.17 Relative recall of the fact extraction for the on-demand created Models, depending on the confidence threshold $\epsilon_{rel}$.

5.18 Excerpt of the model about the relationship between India and Pakistan.

5.19 The first 5 levels of the human cognitive performance category hierarchy. The size of the rectangles indicates the number of descendants.

5.20 Small excerpt of a strongly connected component in the concept graph. For better visualization, classes have been removed.

5.21 Expert scoring of previously unknown facts. The chart shows the fraction of each score as well as accumulated scores for incorrect (1-2), correct (3-6) and both correct and novel (7-9) results.

5.22 Distribution of extraction confidence across the expert scores. High-quality facts get highest scores, but also some incorrect facts were extracted with high confidence.

6.1 Semantic Browser example search.
List of Tables

4.1 Comparison of GlycO to other biomedical ontologies ................. 80

5.1 Example combinations for Focus, Domain and WorldView .............. 103
5.2 Terminology ............................................................................. 104
5.3 Pattern-generalization example .................................................. 116
5.4 Terminology ............................................................................. 116
5.5 Extensional and Intensional similarity between relationships. The top half shows taxonomy relationships specific to the biology domain, the lower half shows domain-independent relationships. The first 2 columns show the relationship pair, the next 4 columns show extensional attributes with the number of shared subject-object pairs and the overall number of instantiating facts for each relationship type. Fraction\textsubscript{min} indicates the fraction of overlap measured by the relationship with the least number of instances. The next 3 columns indicate intensional similarity, computed using pertinence (+\textit{sim}\textsubscript{int}), omitting pertinence (−\textit{sim}\textsubscript{int}) and the difference in similarity. A positive value indicates that +\textit{sim}\textsubscript{int} assigned higher similarity than −\textit{sim}\textsubscript{int} and vice versa. The last column indicates \textit{sim}\textsubscript{rel}, the overall relational similarity, taking extensional and intensional similarity into account. .................................................................................. 123
5.6 Probability that a pattern indicates the \textit{birthPlace} or \textit{occupation} relationship .......................................................... 133
5.7 Comparing the top ten ranked results returned by our method against four competing methods ......................................................... 140
5.8 Sampling of the disjunction of stemmed terms from the glossary and search of Wikipedia for the “mortgage” domain. ............................... 142
5.9 Terms in the Neoplasm subtree of MeSH. Column (1) contains the number for all terms, column (2) contains the number of terms that could also be found on Wikipedia ................................................................. 145
5.10 Parameter Settings for three different experiments to reproduce the MeSH Neoplasms subtree ......................................................... 145
5.11 Criteria, queries and results for the Domain Definition part of Model Creation ................................................................. 156
5.12 Seed-terms excerpt for the Human Cognitive Performance model ............. 156
6.1 Fact browsing statistics. The table shows the assertion that was followed, the number of times it was accessed and its correctness according to domain experts. ................................................................. 173

A.1 All triples that were evaluated by neuroscientists at the Wright Patterson Air Force Research lab. The first column gives the expert score with 1-2 being incorrect, 3-4 correct, but trivial, 5-6 correct general information and 7-9 being correct and not commonly known .................................................... 212
Acknowledgement
My time in graduate school has been an extraordinary journey. It has been fun, exciting, exhausting and sometimes trying, but never boring. It’s a journey that cannot be completed alone and I want to thank all my friends, teachers and collaborators in Athens, GA and Dayton, OH, who shared all or parts of it with me.

First and foremost I want to thank Dr. Sheth, who saw the potential in my research and believed in my abilities. He gave me the opportunity to first work on Knowledge Representation and then switch to Information Extraction. Both together laid the foundations for this dissertation on Knowledge Acquisition. He always created an atmosphere that fostered collaboration, mentoring and personal responsibility. His enthusiasm for innovative research made it possible for me and others in the lab to try and often succeed at thinking big and taking on challenges. His encouragement to do Summer internships allowed me to gain insight into industrial research at Amazon.com and Hewlett-Packard as well as other academic research at the Max Planck institute. It was at HP where I started working with Pankaj Mehra, who was the other great influence in my Ph.D. He first saw the potential of creating semantic models as a Web Service. Working with him taught me how to solve a problem collaboratively and quickly turn ideas into a prototype. Even after the internship, Pankaj tirelessly supported me throughout my research process with deep technical insight and practical suggestions. His involvement was very personal and caring. I am deeply grateful for that.

I am thankful to my dissertation committee members Gerhard Weikum, Shaojun Wang and Pascal Hitzler. I am particularly grateful to Gerhard, who reviewed all of my papers and helped me write with more scientific rigor. At the Max Planck Institute, he gave me an opportunity to conduct my research surrounded by an inspiring group of scientists in my home country.

Amongst my friends I especially want to thank Cartic Ramakrishnan, Pablo Mendes, Karthik Gomadam, Meena Nagarajan, Maryam Panahiazar and Jared Embree. All of them were as deeply involved in my work as I was in theirs and were always there to discuss ideas or read
and critique my papers. Cartic was the one who got me interested in my line of research in the first place and who has been a great friend, roommate and colleague throughout the years. His passion for work, food and life is and has been a constant inspiration. The same holds for Pablo, who also showed me that there is always time to grill a steak Brazilian-style. With Karthik, I bounced many research ideas back and forth to great mutual benefit and great fun while taking our coffee breaks. Meena’s calm and disciplined attitude towards her research work and her communication skills during her Ph.D. set examples for me that I am aspiring to achieve. Maryam’s calm support kept me balanced when I was on the edge of my seat. The final stretch of the journey would have been a lot more bumpy without her. Jared was the most excited about the possible applications of my research and he kept me passionate about it whenever I doubted its practical use. My most loyal friend during the last months of my dissertation work was his dog Sidney, who never left my side and took me out for a run each day.

Most of all I want to thank those who have been there for me from the very beginning. My family, who helped me become the person I am and without whom I would have never had the opportunity to go to graduate school in the first place. My mother, my father and stepfather, my Grandmother and my brother all gave me something on the way that I cherish each day. I will always be deeply grateful for the support that my mother gave me throughout the years, her special brand of loving objective criticism that kept me on my toes and her proverbial seventh sense that made her realize from across the globe that I needed encouragement. My father influenced in many ways how I think about life and work. Even though I did not choose his field of architecture, I still have a desire to see both form and function in my work and make it aesthetically pleasing on a conceptual level.
Dedicated to
my Grandmother
Introduction

All men by nature desire knowledge.

(Aristotle (384 BC - 322 BC), Metaphysics)

Knowledge is arguably the most important factor in advancing human civilization. Despite colloquialisms such as “Ignorance is bliss” there does not seem to be a limit to the amount of knowledge a person, a community, or a system (in a broad sense) should possess. The capacity of our brain may be limited, though, and any person will only be able to hold very small parts of the totality of knowledge held by mankind at any given point in time. This is not only unfortunate, but can be dangerous. We know that in a complex and interconnected world actions can have consequences far beyond the immediate impact. Being disconnected from other fields of inquiry, domain experts cannot see beyond their very own specialized horizons, which become narrower and narrower the more specialized our fields of knowledge become. The silver lining here is that once presented with pertinent information, humans are quite capable of incorporating it into their already held beliefs, adjust their actions and change their world view accordingly. This prompts the necessity of having pertinent information available when needed, so we can expand our knowledge when necessary.
With the largest repository of information ever accessible to us, the World Wide Web is now the primary source of knowledge (McCallum, 2005). Most of this knowledge can be freely searched and retrieved, but very little of it is available formally, in a way that is not only human-readable, but also machine-processable. The last years have seen an onslaught of systems using Information Extraction techniques in order to extract as much information as possible from the Web and formalize the extracted information in machine-accessible form. However, few such systems have made the leap from information extraction to knowledge acquisition.

In this dissertation, I will provide both the theoretical underpinnings and a practical methodology for knowledge acquisition in collaborative systems. I will address this problem from both Knowledge Engineering and Information Extraction perspectives. Unlike traditional approaches, however, this dissertation will show how Information Extraction can be incorporated into a mostly Knowledge-Engineering-based approach, as well as show how an Information-Extraction-based approach can utilize engineered concept repositories.

According to most definitions, only correct statements count as knowledge. This means that information needs to be validated in order to count as knowledge. I will present a semi-automated validation approach that uses deductive methods, as well as a collaborative validation approach that uses social computing.

### 1.1 Motivation

When Sir Tim Berners-Lee laid the foundations for the World Wide Web (WWW) in 1990 by uploading the first HTML page to http://info.cern.ch/hypertext/WWW/TheProject.html\(^1\), a trend was set in motion that revolutionized the way we communicate, consume information and consume goods, but also the way we produce and disseminate data, information and knowledge.

\(^1\)The original URL is not available any longer, but a replica of the page can be found at http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html
Since this first HTML, the Web has been growing exponentially. To date, Google has indexed more than 50 billion pages\(^2\), but because of the dynamic creation of web pages, the potential number of distinct pages is unbounded (At this date, Google uses 1 trillion unique URLs in their Web graph\(^3\)). More importantly, the number of people connected to the Internet has exceeded 34%\(^4\) of the world’s population with a rapidly growing trend, due to increased availability of mobile access through cell phones even in remote areas of the planet and for users who would not be able to afford desktop or laptop computers. The Cisco Visual Networking Index forecast\(^5\) states that by 2015 IP traffic will reach the zettabyte threshold with Internet video taking the lion’s share of the overall traffic. This shows the pervasiveness of the Web and its importance in disseminating data and information, but not yet the impact it can have on creation and dissemination of knowledge.

The Semantic Web (Berners-Lee et al., 2001) is an effort to more meaningfully disseminate data and information by annotating individual information items with grounded concept descriptions (Harnad, 1990). The vision of the Semantic Web, however, is more ambitious. It aims at being a Web of Knowledge that contains formal descriptions of the world in the form of so-called ontologies\(^6\). The term ontology is borrowed from philosophy, where the discipline of (upper case) Ontology is concerned with the inquiry into the nature of being and categories of existence. In contrast, an ontology (lower case) in information systems jargon is a formal account of the concepts and their inter-relationships pertaining to a domain of interest. Whereas Ontology tries to analyze the nature of things, ontologies normatively prescribe the nature of things as they pertain to a domain or a task.

Much of the research literature on the Semantic Web ascribes the task of knowledge acquisition to either individuals, groups of people, or automated mechanisms. This is reflected in the fields of ontology design and knowledge representation (Sowa and Others,\(^2\)\(^3\)\(^4\)\(^5\)\(^6\))

\(^2\)http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
\(^3\)http://www.worldwidewebsize.com/
\(^4\)http://www.internetworldstats.com/stats.htm
\(^6\)The use of the plural form of the word ontology is technical jargon, and not a construct of philosophy.
2000; Guarino, 2009) as well as ontology learning (Maedche and Staab, 2001) and Information Extraction (Sarawagi, 2008). However, despite the great attention that has been given to these aspects of knowledge acquisition and information extraction, little attention has been paid to the particular challenges of knowledge acquisition in a large collaborative and interactive system. That is, the Semantic Web has not fully considered the Web as a knowledge-producing entity.

1.2 Hypothesis

The World Wide Web gives us the ability to connect mankind within a large system. People and machines, i.e. human and artificial agents interact synergistically to increase the amount and availability of knowledge within the system. For the purpose of this dissertation, the Web is seen as a collaborative environment shared between humans and machines. The user is thus conceptually an integral part of the Web. With this premise, we can view the Web as a closed system; the interaction with the world outside the Web is accomplished by the users who act as an interface. Because of this interface, the Web itself is sufficient for acquiring new knowledge about the World both inside and outside this system. Information published to the Web is analyzed, potential knowledge extracted from it and delivered back to the system for verification. At the end, only statements that can be derived through logical deduction or statements that have been vetted by a community of users should count as actual knowledge. This view allows us to produce formal knowledge in a self-perpetuating hermeneutic fashion by extracting and verifying the extracted information within the system. Figure 1.1 illustrates this cycle of knowledge acquisition.

This dissertation outlines a framework for knowledge acquisition within a system. To show that the techniques that allow efficient knowledge acquisition change depending on the properties of the system, I will structure the dissertation around two examples: (1) Formal Knowledge acquisition in a small, tightly connected system, such as a laboratory, a
research center or a research community and (2) Formal Knowledge acquisition in a large, loosely-coupled system, such as the Web. I will explain epistemological foundations for the acquisition of knowledge and subsequently show implementations of this framework that use (for 1) Knowledge Engineering techniques coupled with Information Extraction techniques and (for 2) Information Extracted from formal and informal sources in conjunction with a social-computing-approach to validation of the extracted information.

Information is most useful when it is packaged as a coherent set of assertions that are pertinent to the information seeker’s interest and intentions. Ideally, the information comes in a comprehensive model that exactly describes the concepts and relationships as well as the facts that instantiate the relationships that are important in context of the domain of interest.

In scientific domains this prompts the necessity of having commonly held explicit and tacit (implicit) knowledge formally available. Scientists should not spend time searching for fairly well-understood concepts in their own field or in adjacent fields. However, due
to specialization of disciplines, what is common knowledge in one field is often either unknown or only superficially known by scientists in other fields, sometimes even when the fields are relatively close. Here it is particularly important to have a formal account of even the basic concepts of a field, because it is often a lack of understanding of basic principles that critically underlies the outsiders inability to comprehend a research paper, for instance. Arguably, models used in such interdisciplinary contexts must be created and evaluated by field experts to only contain valid, agreed-upon statements. Careful knowledge engineering is required to create such formal models of scientific value. In Chapter 4, I will show how this can be achieved while still using automated techniques to extract specific entities of interest for the domain. I will explain how ontology design can be studied within the cyclic framework of knowledge acquisition, albeit by applying the framework to a small, tightly connected system.

Outside the scientific realm, concise knowledge models are often lacking. Focused information about new topics needs to be generated on demand. Breadth of knowledge is more important than depth, and the outsiders perspective matters for the model’s relevance. This dissertation will describe a system for on-demand domain model creation from a domain outsiders perspective. Such models represent domain knowledge as formal statements, either in first-order logics or in graph formats such as OWL (Horrocks et al., 2003) or RDF (Manola and Miller, 2004). The model contains concepts and relationships, as well as entities and facts, pertaining to the domain of interest. Automatic creation of such models needs to build upon any existing formal knowledge, as well as needs to extract new statements from informal data in order to extend that knowledge. In the broader context of knowledge acquisition, this approach can be viewed as socially driven knowledge acquisition, because more models will be created – hence more knowledge made formally accessible – for topics of mutual interest to many people.

Knowledge acquisition in the context of the World Wide Web is therefore naturally viewed as a cyclical process involving available background knowledge, information ex-
traction and validation of extracted assertions through Web communities. The circle is completed when validated acquired knowledge becomes part of the common background knowledge.

The models created by our framework can be used to aid in search, browsing and content classification. A domain model here is a representation of a field of interest that does not quite target the representational rigor expected from formal ontologies, but nevertheless provides a concise and closed conceptual description of a domain with categories, instances, and relationships. Since the models are created on demand, high-interest domains will soon also have a stronger representation on the Linked open Data (LoD) cloud (Bizer et al., 2009a) after the extracted facts have been vetted. Very often the process of evaluating the model, i.e. evaluating the extracted facts in the model, is similar to the intended use of the model, for example in guided browsing or classification.

The technical centerpiece of automated knowledge acquisition in this dissertation is the automatic creation of semantic models for focused domains of user interest. The reason why it is of major importance is that a) the amount of formal knowledge on the Web will not grow significantly without automated procedures and b) using automated methods to help individual users formalize their domains of interest will lead to participation in creation and validation of information.

The models will be created in a two-step process, with each step drawing from different kinds of information sources. The first step is the automatic creation of a domain hierarchy that sets the boundaries for the area of interest. This hierarchy is extracted from Wikipedia whose semi-structured corpus supports statistical inferences about the interrelatedness of its concepts, based on its internal link structure and category structure. Moreover, Wikipedia concepts are unique and well-defined.

The second step uses an Information Extraction algorithm to automatically relate pairs of Wikipedia concepts by learning from the linguistic patterns used for textually instantiating the various relation types in background knowledge sources, such as LoD or the UMLS
metathesaurus.

To clarify the differences between Information Extraction (IE) and knowledge acquisition, one has to go to the definitions of information and knowledge. In the context of information systems, the DIKW hierarchy has often been cited as giving a good distinction between data, information, knowledge and wisdom (Rowley, 2007; Ackoff, 1989). This hierarchy is also referred to as the “Information Hierarchy” or the “Knowledge Pyramid”. It is described as follows:

- **Data** are defined as symbols that represent properties of objects, events and their environment. They are the products of observation. But are of no use until they are in a usable (i.e. relevant) form. The difference between data and information is functional, not structural.

- **Information** is contained in descriptions, answers to questions that begin with such words as who, what, when and how many. Information systems generate, store, retrieve and process data. Information is inferred from data.

- **Knowledge** is know-how, and is what makes possible the transformation of information into instructions. Knowledge can be obtained either by transmission from another who has it, by instruction, or by extracting it from experience.

- **Wisdom** is the ability to increase effectiveness. Wisdom adds value, which requires the mental function that we call judgment. The ethical and aesthetic values that this implies are inherent to the actor and are unique and personal.

  (Rowley (2007))

For this procedural definition of knowledge, the distinction between information and knowledge is of primary interest. In the case of this hierarchy, it is the distinction between
“Know-what” and “Know-how”, i.e. the difference between descriptions and instructions. Epistemologically, however, the concept of knowledge goes beyond the notion of holding facts in mind or holding facts in a formal representation. Knowledge is often defined as “justified true belief,” which means that in order to know something, we have to believe it, we have to be justified in believing it and it needs to be true. A system may not be capable to experience the state of mind of believing, but the metaphor holds when we accept that every statement that is formally available in a system is held with some degree of belief, be that implicitly or explicitly. Justification can be given by using a valid method to acquire information on the one hand and a valid method to validate the information on the other. The biggest difficulty is presented with the truth requirement. Unless a statement can be deductively derived from some axioms, the truth of the statement may not be accessible.

Requiring knowledge to meet the epistemological requirements is important. The procedural definition does not necessarily require statements to be true and may be better described as “actionable information.”

The aim of the dissertation is to show one possible approach to automated, nearly perpetual knowledge acquisition within a system. This is contrasted with a traditional knowledge engineering approach. Both serve different purposes, apparent not only in their applications but also in the structures of the resulting knowledge representations.

The hypothesis of this work is that formal knowledge can be generated in a large system in a perpetual manner, given that mechanisms are available for representation, extraction, and validation of knowledge. The crucial validation step from extracted information to knowledge is required to be performed by a justification procedure that, in most cases, needs to involve humans. For this reason it is beneficial to have a user-driven extraction of information in order for the users to be interested in validating the extracted information. Recent research has shown that users are willing to engage in productive activities on the Web if certain conditions can met, e.g. if it is fun to help (Von Ahn and Reiter, 2005), if it is beneficial for the user monetarily if it serves a greater purpose (Thomas and Sheth,
Wikipedia is a shining example for communal creation of correct and validated information. In this dissertation it is assumed that the creation and validation of knowledge can be done in a more formal and more distributed way than on Wikipedia, where a) the creation and validation is done within a single environment and b) the creators of the information usually need to be sufficiently committed to the topic of interest to invest large amounts of time. With the extensive amount of information present on the Web and sophisticated information extraction techniques, it is conjectured that users can assist in the creation of knowledge to a large extent by simply using the information that was extracted and by investing little or no extra time.

Domain model creation was chosen as the focal point for knowledge acquisition, because the assumption is that knowledge is best created in context, with specific goals in mind. Users that have an individual interest in acquiring knowledge of a specific domain are more likely to engage in an active discourse with a system to allow for validation of information. This is important, because no automatic information extraction approach will ever be 100% accurate. Any extracted assertion, unless it can be deduced from previously validated statements, necessarily needs human validation in order to be called knowledge. That is not to say that humans will always be correct.

The intuitive approach to validating assertions is by letting experts vote on their correctness. However, given that vast amounts of statements can be automatically extracted, I propose a Validation through Use approach, in which assertions are implicitly validated by using them in Information Retrieval, Classification, Search and Browsing tasks. The user behavior when interacting with the extracted assertions or with the information provided on the basis of the extracted information gives us an idea of which assertions may be correct, which may be incorrect. A focused browsing application of this paradigm will show different ways to use and evaluate the extracted information. Once validated, mere assertions can be seen as facts and added to a larger pool of knowledge, such as the LoD cloud.

A corollary of the knowledge acquisition hypothesis brought forward here is that eval-
uation is an integral part of the knowledge life cycle. Most theories of knowledge life cycles implicitly make this assumption, but it has not been thought of in a systemic fashion. When we start from using socially created information and knowledge as the basis for information extraction, it is but a small step to also use human/social computation in order to verify facts and to add new knowledge to existing background knowledge. If we think of this background knowledge on a Web-scale, for example in LoD, this approach can help advance the overall state of knowledge on the Web.

1.3 Scope

This dissertation will outline a theory and practice of knowledge acquisition in collaborative environments or systems. It will juxtapose two different styles of knowledge acquisition that are each appropriate in different types of systems. One Knowledge Engineering centric approach that can be most successful in a smaller, tightly coupled environment and one Information extraction centric approach that is suitable within the scope of a large system such as the World Wide Web.

Chapter 2 gives a summary of the conceptual and technical contributions of the dissertation. Chapter 3 introduces epistemological concepts, namely knowledge, justification, belief, and truth as well as describes conceptual differences in the acquisition of knowledge between individuals, groups and systems. Chapter 4 describes the design of an ontology in the biochemical domain. It demonstrates the need for complex representations in specialized domains and introduces domain-dependent ways to automatically increase the amount of knowledge, juxtaposing this chapter with the domain-independent knowledge acquisition in the following chapters. Chapter 5 describes efforts to automatically create taxonomies or hierarchies of domains of interest by connecting concepts in a hierarchy with semantic relations. Knowing that automatic extraction does not produce infallible results, Chapter 6 shows how to verify the automatically extracted information using community efforts and human computation. Chapter 7 finally concludes the Dissertation.
This dissertation takes on some issues discussed in Welty and Murdock (2006). In their discussion of the integration of Knowledge Engineering and Information Extraction the authors identify five dimensions of interoperability problems:

1. Precision
2. Recall
3. Relationships
4. Annotation vs. Entities
5. Scalability

The discussion brings up some important points. Precision in IE is to date never reliably perfect, which means that for a Knowledge Acquisition task, IE techniques by themselves are not adequate. Recall is problematic, because in most cases, a human reader would get more information from a document than any generic automated method can. However, the assumption in this dissertation is that conceptual knowledge is not extracted from a single document, but from the entire available document collection. Extraction that
aims at getting most information out of a single document should rather be seen in the context of text summarization than knowledge acquisition. The hypothesis of this dissertation also states that a single mention of a statement is not a reliable evidence for its factual character.

Relationship (i.e. n-ary predicates with \( n \geq 2 \)) are harder to extract than type/class assignment, which are unary predicates. In free text, however, unary predicates are usually expressed in the same general ways as binary predicates. We seem to think about a type relationship as \( \text{is}_a(\text{entity}, \text{tyoe}) \), rather than \( \text{type}(\text{entity}) \), for example \( \text{is}_a(\text{Opus, Penguin}) \) instead of \( \text{Penguin(Opus)} \).

The issue of annotation vs. entities brings up an important point, even though I believe that the term ‘annotation’ is poorly chosen and should rather say ‘entity mention’ or ‘entity reference’, because the entity mention does not present a remark to the entity, but rather refers to it. The point remains, however, that text only refers to entities, it does not “contain” them. Even disambiguation and coreference resolution by themselves do not solve this problem, because there is still no grounding of the concepts that are mentioned in the text. This dissertation therefore starts with a grounded representation of concepts and attempts to find the referring concept mentions in text. The issue of coreference resolution is sidestepped by assuming that in a large-enough corpus, facts will be expressed multiple times, possibly with different concept designators/labels for different occurrences. It is then the collection of factual expressions, rather than a single occurrence, that gives the algorithms confidence in the formal statements that are extracted from these expressions.

The interoperability that is highlighted by Welty and Murdock (2006) brings about a question of how much of each technique should be used in an application. I will address these issues from two different directions, exemplified by two Knowledge Acquisition projects; one that is more Knowledge-Engineering oriented, one that is more Information Extraction oriented (See Figure 2.1).
Figure 2.1: Classification of the work in this dissertation in terms of Knowledge Engineering vs. Information Extraction

The circle of knowledge as an abstract guideline for Knowledge Acquisition assumes a current state of knowledge that is then improved by learning and validating new knowledge items, i.e. concepts and facts involving both the learned as well as the already known concepts. Before the validation phase, facts are merely statements that are waiting to be verified. This dissertation offers two approaches to knowledge acquisition within this framework. The differences are in the assumptions that in some cases, background knowledge is mostly tacit and needs to be formalized by domain experts and knowledge engineers in a discourse, whereas in other cases, the background knowledge needed is available in the form of formal statements. The learning phase is in either case split into the acquisition of a Domain Definition and a Domain Description. This reflects the idea that a) the presence and the definition of concepts is usually less contentious than their properties and b) formal concept designators are more abundantly available and can more easily be automatically extracted. In a highly axiomatized system, validation can be done deductively using automated reasoning techniques, whereas in systems that lack a clear axiomatic underpinning, validation requires extensive human involvement.

Depending on the application of the knowledge, it is acceptable to have a shallower representation or a more in-depth representation of the domain knowledge. Ontologies or
domain models for Information Retrieval applications do not need to be highly expressive in order to be useful. Rather, they should contain a fair number of entities that are important for a domain and ideally have alternative terms/synonyms for the entities and concepts. In applications that require reasoning over data, it is important to have a highly expressive ontology that is able to deduce or to be queried for complex relationships between concepts or entities.

The work on *GlycO* (Chapter 4) is an example of the latter. It mostly uses Knowledge Engineering techniques methodologies to create a highly expressive T-Box as well as a set of so-called “Archetypal Instances” that encode domain experts’ knowledge about complex carbohydrate structures and interactions. However, it also has an information extraction part that uses the knowledge in the Archetypal Instance to extract formal descriptions of complex carbohydrate structures from text or carbohydrate databases. The archetypal instances encode tacit domain knowledge in addition to explicit knowledge, i.e. properties that a real world structure that corresponds to the instance has and that is well known, but rarely expressed by experts is made explicit in the formal description. Chapter 4 will describe this interoperation of strong KE and weak IE in-depth.

The work on *Doozer/Doozer++* (Chapter 5) is an example of the former. Even though its mechanism to create ontologies or domain models is fully automated, it takes advantage of weak Knowledge Engineering techniques in the form of socially constructed knowledge bases, such as Wikipedia and DBPedia (Bizer et al., 2009b).

### 2.1 Terminology

In this work the terminology used will mostly be familiar, however, I will define some terms here to ensure a consistent understanding.
Terms denoting the work that underlies this dissertation

**GlycO** Complex Carbohydrate (Glycan) domain ontology.

**Doozer** Domain Hierarchy creation application.

**Doozer++** Current evolution of Doozer, including fact extraction.

Technical terminology used throughout this dissertation

**Concept** Generally defined as a unit of knowledge. Here it refers to an individual, a class or a property type in an ontology or a domain model.

**Term** Single noun or compound noun phrase that denotes a concept.

**Entity** A thing with a separate, self-contained existence. Here it refers an individual in an ontology or a domain model.

**Class** An entity type.

**Category** Structure on Wikipedia to organize articles into a hierarchy.

**Statement** An assertion of a relationship involving concepts.

**Fact** Actual state of affairs, manifested in a validated statement.

**Pattern** A set of things (events, objects, etc.) that occur and repeat in a predictable manner. In this dissertation, a pattern is always a sequence of textual tokens that represent a binary relationship.
2.2 Knowledge Engineering oriented knowledge acquisition

There is a large body of work in the area of Knowledge Engineering and ontology design both from domain independent and from domain-specific points of view, which is mostly concerned with the acquisition of T-Box knowledge. This dissertation will not add to the discussion in this area. Instead, it will show a way to automatically add highly refined A-Box knowledge in the presence of a well-formed T-Box and a domain-specific entity extraction algorithm.

Developing a highly expressive formal ontology for a comparatively narrow field of research requires the constant interaction between domain experts and knowledge engineers. The modeling of knowledge calls for a profound understanding of a domain. The domain expert must fully participate in ontology development and understand the formalisms used for specifying the conceptualization of the domain. Conversely, the knowledge engineer must analyze the ontology for formal consistency and semantic correctness to avoid ontological fallacies in modeling. The Ontoclean methodology (Guarino and Welty, 2002) explains how concepts should be classified on a meta-level according to distinctions like rigid versus non-rigid concepts, entities versus roles, etc. The knowledge engineer must have enough domain knowledge to apply these distinctions to the ontology.

When designing complex domain ontologies in highly specialized fields, it is not only necessary to come to a common understanding, i.e. for domain experts and knowledge engineers to find a shared language, it is usually up to the knowledge engineer to elicit the tacit domain knowledge from the expert (Johnson, 1983; Nonaka and von Krogh, 2009) and then put it in a formal language.

As individuals master more and more knowledge in order to do a task efficiently as well as accurately, they also lose awareness of what they know. The very knowledge we wish to represent in a computer program as well as
the knowledge we wish to teach others, often turns out to be knowledge that
individuals are least able to talk about. (Johnson (1983))

In the development of GlycO, the parts of Domain Definition, Domain Description and Validation were accomplished as follows:

**Domain Definition** The T-Box of the ontology was manually created by engaging in a discourse with the domain experts and iteratively closing the gap between the domain conceptualization that is in the experts’ minds and the formal representation thereof. The separation between Domain Definition and Description is thereby mostly conceptual, as some descriptions and restrictions on concepts are asserted in this phase already.

**Domain Description** Given a comprehensive domain definition and a highly axiomatized description of central concepts, this knowledge can be used to create domain-specific extraction algorithms that can achieve a very high precision. In this case, extracted concept descriptions can be augmented with knowledge that was not necessarily present in the extraction source, but is the tacit knowledge that was formalized during the definition phase.

**Validation** Given a clear domain axiomatization and a rich description of extracted statements, a reasoner can identify whether a statement is entailed by the knowledge base. A successful proof of entailment guarantees the correctness of the statement in this context. However, if the statement is not entailed the proof possibly failed because of an incomplete knowledge base, rather than in incorrect statement. In this case, the statement is given to a panel of experts for validation.
2.3 Information Extraction oriented knowledge acquisition

One reason why general-purpose IE approaches to Knowledge Acquisition have not received more attention despite the apparent attractiveness of automated methods is that IE is still very unreliable. Knowledge Acquisition poses strong requirements on the mechanisms used to extract and on the veracity of the statements that are extracted. Nevertheless, the vast amount of information that is available on the World Wide Web makes this a worthwhile endeavor.

The framework towards perpetual knowledge acquisition involving automated domain model extraction relies on the following factors and components:

1. Background knowledge
2. Free textual information
3. Information extraction techniques
4. Knowledge validation mechanisms
5. Knowledge management to merge new knowledge

It is assumed that background knowledge is available in the form of ontologies or Linked open Data (LoD). It is also assumed that information is freely available on the Web. This dissertation will not discuss the quality of information or the merging of knowledge after being extracted. However, there is a vast amount of work on information quality available, see for example (Rieh, 2002; Knight and Burn, 2005). My own work on the topic is with regards to the quality and maturity of articles on Wikipedia (Thomas and Sheth, 2007). Merging of formal knowledge has also been widely discussed in the field of ontology alignment. Here see for example (Euzenat and Shvaiko, 2007). My own work
on the topic uses Expectation Maximization to find the best matches between ontologies (Doshi et al., 2009; Doshi and Thomas, 2006).

The focus here is on the extraction of knowledge in the form of formal context models that describe a user’s domain of interest with respect to e.g. a web or news query or a classification task. The approach taken here builds focused domain models by a top-down extraction of a conceptual lattice from Wikipedia and then uses bottom-up Information Extraction techniques to find semantic relationships between the domain concepts. The resulting context/domain model is then validated by either the user who created it or by social computing techniques using implicit and explicit validation methods.

2.3.1 Epistemological Considerations

In the introduction I gave a functional definition of knowledge in the context of the DIKW hierarchy (Rowley, 2007; Ackoff, 1989). This definition was as follows:

Knowledge is know-how, and is what makes possible the transformation of information into instructions. Knowledge can be obtained either by transmission from another who has it, by instruction, or by extracting it from experience. (Rowley (2007))

Other functional definitions from the more recent literature go in the same direction. Bird (2010), for example, sees knowledge as an input to deliberation and action. In these functional views, knowledge usually guides and enables action (Hawthorne and Stanley, 2008).

A functional view of knowledge becomes important in the application of the knowledge models that are the focus of this work. However, this view also fits the term actionable information and may thus encompass pieces of information that can not formally be called knowledge according to more epistemologically rigid definitions. For example, the above definition does not require a statement to be correct in order to qualify as knowledge. In a normative sense, actions and instructions should be based on correct information that fits a
narrower definition of knowledge. Traditionally, knowledge has been defined as “Justified true belief”. Despite some challenges to this tripartite definition, it can well serve as a basis for a discussion. When taking this as a minimum requirement for the production of knowledge, it is important to see which parts of the technical framework meet the categorical requirements given by the definition.

In social epistemology, it is believed that there is a general effort towards making truthful statements (Goldman, 1999). This means that we can generally trust a majority of the statements we find in publications and on the web. However, it is clear that some statements will be incorrect and need to be identified or at least disregarded. The Wisdom-of-the-crowds paradigm states that a crowd will in the long run give the best answers to a question, given the the individual answers are aggregated is sound and takes potential outliers into account (Surowiecki, 2005; Thomas and Sheth, 2011).

Applying this maxim to Information Extraction tasks, an assertion should only be truster when it is asserted by various sources or when the source is created in a collaborative or peer-reviewed manner. This assures that the extracted domain models, despite their focus on an individual’s interest, represent a shared understanding of the knowledge that underlies the model creation, thus fulfilling one of Gruber’s requirements for formal ontologies (Gruber, 1993a). Moreover, there is usually general agreement about the existence of a concept, an event, an entity. The disagreement tends to appear with the description of the concept and by relating it to other concepts. Hence, it is a safe assumption to make that initial concept designators can be harvested from existing encyclopedic sources, whereas relations should be extracted from free text in an aggregating manner.

From the perspective of analytical philosophy, an ontology can be seen as a combination of rigid concept designators (Kripke, 1980) in the form of URIs together with definite descriptions (Russell, 1905) in the form of assertions about these concepts. Along these lines the task of domain model creation is split into the two parts of Domain Definition and Domain Description.
Domain Definition builds a conceptual basis for the model, expressing what entities exists and by which types of relationships they can generally be related, but not how they are actually related. The term “concept” is used here in a general definition as a “unit of knowledge” that comprises classes, individuals or relationship/property types in an ontology. Domain definition is thus similar to, but not the same as the T-Box, because it includes individuals, whereas the T-Box of a domain model only contains classes, properties and their logical restrictions. The Domain Description contains facts, i.e. instantiations of properties that were defined during the definition step.

The manner in which the concepts are gathered in the domain definition assure that there is a real-world correspondence with the model. It also assures that the model actually contains concepts rather than just concept mentions, which is a common criticism of IE techniques for Knowledge Acquisition (Welty and Murdock, 2006). For the conceptualization, it thus also meets the need for justification of the presence of the concept.

The requirements for truth and justification are more difficult to meet in the case of extracted factual assertions. This is referred to as the domain description. Whereas there is less doubt about the existence of a concept, its attributes are usually more contentious. The method used for domain description in this work is a concept-centric pattern-based IE algorithm that gathers evidence for the existence of a relationship between concepts from multiple sources. This assures a sensible justification on the algorithmic side. The final justification and confirmation of the correctness or truthfulness of an extracted formal statement must however be given by a community of peers. This is accomplished by social computing methods for the validation of statements described in chapter 6.

2.3.2 Automatic Domain Model Creation

The conceptual separation of Domain Definition and Domain Description is reflected in the use of different techniques for both tasks. If the domain is defined by extracting known concepts from a corpus that already contains clearly delineated concepts with unique identi-
fiers, ambiguities can be avoided and the description step can work on the basis of concepts, rather than just terms. Conversely, in many NLP-based approaches, ontology learning is based on promoting phrases to predicates. In these cases, no clear denotation and no clear designator are given, because only concept mentions are extracted. Thus the statements in the ontology often fail to refer to actual entities, events or states of affair. If we have information sources that give us an idea of the identity of concepts and specific types of relationships, we know that the extracted concepts and relations do actually refer.

To put the idea of combining top-down and bottom-up approaches in context, I will show a main difference to previous approaches. Buitelaar et al. (2005) present an ontology learning layer cake with terms on the bottom and rules on top (Figure 2.2). The layer cake suggests that the learner goes from terms to synonyms to concepts and concept hierarchies, before adding relationships and rules. This approach can and should be taken when the only information available is in the form of raw text. However, the steps from terms and synonyms to concepts are error-prone when done automatically. When conceptual knowledge is available in the form of taxonomies, encyclopedias or thesauri, this knowledge should be harvested. Humans are much better at intuitively identifying concepts than machines are and this capability is reflected in these knowledge sources. For this reason, this approach starts with concepts rather than terms. Instead of extracting previously unknown concepts from text, concepts are assumed to be available and their existence needs to merely be verified in the text corpus. Once evidence is found in the text corpus for the existence of the concept further knowledge about the concepts can be gained through descriptions in the form of relationships (see figure 2.3).

To summarize the conceptual considerations, this work is built on the following premises:

1. Humans succeed at identifying, defining and describing concepts

2. Ontologies represent a human conceptualization and abstraction of the world
3. Unambiguous (semi-)formal concept designators and identifiers are available in greater abundance than concept descriptors and relations.

   (a) Encyclopedias, glossaries and vocabularies provide concept designators

   (b) Community-created or peer-reviewed Encyclopedias, glossaries and vocabularies express a shared view of a domain

⇒ Extract domain definition top-down from such corpora

4. Concept descriptions such as attributes and relationships are plentiful in informal text.

5. Concept descriptions are manifested in multiple documents.
6. An aggregation of multiple statements about a concept yield a more accurate description

(a) A macro-reading-based (Mitchell et al., 2009) Information Extraction approach aggregates distributed information

(b) Pattern-based IE inherently conforms to an aggregative process

⇒ Extract/improve domain description bottom-up from free text

**Domain Definition**  A focused set of concepts that define the scope of a domain provides a grounding and a contextualization of the knowledge acquisition task. *Domain Definition* is accomplished by restricting existing structured or semi-structured sources to only contain concepts pertinent to a focus domain. In this work, Wikipedia is used as a knowledge source. under the assumption that most concepts and entities that most users are interested in are represented by articles in Wikipedia. *Domain Definition* can be compared to engineering a T-Box of an ontology, which specifies the types of concepts and relationships/attributes that can exist in a domain. However, a T-Box does usually not contain definitions of entities, whereas the *Domain Definition* incorporates entities that are of interest in the domain.

**Domain Description**  According to the procedural definition of knowledge, the step from information to knowledge is from the “know-what” to the “know-how”. With a domain definition, it is already known which concepts are of interest. In this next step, the “know-how” is acquired by finding facts involving the domain concepts in order to have a description of the concept interactions and dependencies. Facts put the mere definitions of concepts into perspective by relating them to other concepts or endowing them with attributes. By definition, facts are verified statements that refer to actual states of affairs. Hence it is important to have a measure of confidence as well as rigorous testing of extracted statements in order to assure correctness.
**Evaluation and Validation in Use**  The idea of extrinsic evaluation or evaluating an algorithm in use, i.e. analyzing the user’s interaction with the software, has been of interest mostly in the design community. Bannon (1996) realized early that evaluation should be an integral part of an application. His work is concerned with design applications, but the principle applies to IR, IE or Knowledge Acquisition as well.

My goal is to integrate the validation seamlessly into the application of the knowledge. In IR research, great strides have been made, notably by Agichtein et al. (2006a) and Bian et al. (2008), to tune Web search ranking to selection preferences. However, finding more relevant results closer to the top of search results is mostly an issue of convenience rather than of absolute correctness. For Knowledge Acquisition, the stakes are higher in terms of the required accuracy of facts that a user interaction should provide. To be absolutely certain about the correctness of a fact, it has to be given directly to a panel of experts that approve or disapprove. Given the amount of assertions that can be extracted using automated methods, though, this is not feasible for all facts. User interaction with extracted statements can point to those that are most likely correct. In Chapter 6 I will discuss the methodology behind the envisioned “Validation in Use”.

We never are definitely right, we can only be sure we are wrong.

(Richard Feynman (1967), The Character of Physical Law, p. 152.)

3.1 Introduction

This chapter outlines an epistemological background to knowledge representation, acquisition and propagation as it pertains to this dissertation. It will discuss different definitions of the concept Knowledge and will justify the hermeneutic approach to knowledge acquisition that the dissertation is built upon. Further, it will discuss the change in the concept of knowledge when it is seen in light of a system rather than an individual, i.e. the notions of subjective and objective knowledge.

In the introduction I gave a functional definition of knowledge in the context of the DIKW hierarchy (Rowley, 2007; Ackoff, 1989). This definition was as follows:

Knowledge is know-how, and is what makes possible the transformation of information into instructions. Knowledge can be obtained either by trans-
mission from another who has it, by instruction, or by extracting it from experience. (Rowley (2007))

Other functional definitions from the more recent literature go in the same direction. Bird (2010), for example, sees knowledge as an input to deliberation and action. In these functional views, knowledge usually guides and enables action (Hawthorne and Stanley, 2008).

A functional view of knowledge becomes important in the application of the knowledge models that are the focus of this work. However, this view also fits the term actionable information and may thus encompass pieces of information that can not formally be called knowledge according to more epistemologically rigid definitions. For example, the above definition does not require a statement to be correct in order to qualify as knowledge. In a normative sense, actions and instructions should be based on correct information that fits a narrower definition of knowledge. In the following section I will give a categorical definition that identifies knowledge as “Justified true belief” and thus implies correctness.

3.2 Knowledge

Knowledge has traditionally been defined as “Justified true belief” (Chisholm, 1982; Plato and Campbell, 1883). This means that an assertion, in order to be called knowledge, needs to be believed by the entity holding the knowledge, it needs to be justified by some mechanism and it needs to be true.

This definition is usually presented in the following tripartite approach:

A subject S knows that a proposition P is true if, and only if:

1. P is true

2. S believes that P is true, and

3. S is justified in believing that P is true
This tripartite definition can be seen as reflecting a universal component (Truth), a social component (Justification) and an individual component (Belief). Thereby the following considerations are important:

- **Belief**: Describing belief as the “individual’ component’ should hereby not be understood as necessarily belonging to a single person. A system or a social group can be seen as an individual entity that holds a belief.

- **Justification**: If the justification procedure is not a social mechanism by itself, such as a discussion or a vote, it is usually approved by such a social mechanism. For example, it has at some point been agreed upon that induction is a good justification procedure, even though it may lead to false conclusions. The same holds for witness testimony in courts.

- **Truth**: Seeing truth as the universal concept of the three may be the most contentious proposition. I will discuss different theories of truth below to highlight the difficulties with the concept of truth.

Taking the social aspect of justification into account, the acquisition of knowledge is by definition a social endeavor. An unjustified belief is not knowledge, whether it is true or not. It can merely be called a (more or less educated) guess (Olen, 1976). The winner of a large lottery jackpot, for example, did not “know” the lottery numbers, but made a lucky guess. Hence it is important to have a good reason for believing an assertion for it to constitute knowledge.

The truth condition may seem counterintuitive at first, because we claim to “know” so many wrong things. However, normatively, the wrong things we seem to know are merely false beliefs and we would likely not attribute the knowledge of false statements to others. For example, we would not think that anybody “knows” that the earth is flat. The belief condition is important, because it necessarily precedes knowing. Knowing is seen as a state
of mind or a state of a system. If I do not believe that the earth revolves around the sun, I do not have that knowledge, even though it is true and there are good reasons to believe it.

Justification, truth and belief are thus seen as necessary conditions for a statement to constitute knowledge. A major challenge to the tripartite definition of knowledge came from the so-called Gettier problems (Gettier, 1963), in which Gettier shows that the three conditions are not always sufficient for an assertion to constitute knowledge. Gettier’s constructed examples, even though far-fetched and not widely applicable, show that an individual that holds a “Justified True Belief” can still not know. In these cases the statement holds true accidentally, even though the intended reference of the statement is broken.

For clarification, see Gettier’s first example from his paper:

Suppose that Smith and Jones have applied for a certain job. And suppose that Smith has strong evidence for the following conjunctive proposition:

(d) Jones is the man who will get the job, and Jones has ten coins in his pocket.

Smith’s evidence for (d) might be that the president of the company assured him that Jones would in the end be selected, and that he, Smith, had counted the coins in Jones’s pocket ten minutes ago. Proposition (d) entails:

(e) The man who will get the job has ten coins in his pocket.

Let us suppose that Smith sees the entailment from (d) to (e), and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true.

But imagine, further, that unknown to Smith, he himself, not Jones, will get the job. And, also, unknown to Smith, he himself has ten coins in his pocket. Proposition (e) is then true, though proposition (d), from which Smith inferred (e), is false. In our example, then, all of the following are true: (i) (e) is true, (ii) Smith believes that (e) is true, and (iii) Smith is justified in believing that (e) is true. But it is equally clear that Smith does not know that (e) is true; for
(e) is true in virtue of the number of coins in Smith’s pocket, while Smith does not know how many coins are in Smith’s pocket, and bases his belief in (e) on a count of the coins in Jones’s pocket, whom he falsely believes to be the man who will get the job.  

(Gettier (1963))

Most Gettier-type problems aim at showing that the criteria we are using to call an assertion knowledge are too broad and allow for assertions to slip in that are not knowledge. More formally, as stated by Kirkham (1984):

A Gettier type counterexample is used to show that a proposed analysis of knowledge is too inclusive. Such counterexamples are hypothetical situations in which (1) all of the conditions for knowledge specified in the analysis are met, but (2) the epistemic agent does not have knowledge because the conditions have been met only by dumb luck, by accident, by coincidence, or by some means we intuitively regard as illegitimate.  

(Kirkham (1984))

In other words, there is “a lack of successful coordination between the truth of \( p \) and the reasons that justify \( S \) in holding that \( p \)” (Floridi, 2004). To rule out this kind of coincidental knowledge, it is often argued that another condition should be added to assure that knowledge is consciously held. Lehrer and Paxson (1969), for example, adds the requirement of indefeasibility of a statement to circumvent Gettier problems. However, Floridi, however, argues that Gettier problems are unavoidable in the tripartite approach, even if we try to strengthen any of the conditions or add more conditions.

If Floridi (2004) is correct in stating that gettierization cannot be avoided in the tripartite approach, we are left with two options: a) we could dismiss the approach and redefine knowledge or b) accept that a few statements may be false positives. I believe that while a) should be pursued, b) is the better option for the remainder of this dissertation, because the tripartite character of knowledge and its interpretation as having an individual, a social and a universal component fit very well with my goal of knowledge acquisition in a collaborative system.
Even after accepting that knowledge is “justified true belief”, the definitions of the individual concepts are contentious. This dissertation will not aim at finding new definitions that are more applicable to acquiring knowledge in a system, but rather show how using the traditional definitions can yield an epistemologically justified knowledge acquisition. I will give a broad overview over the different approaches at defining them.

3.2.1 Truth

The definition of truth ranges from strict normative theories that make attaining a true statement almost impossible to relativistic theories that put the concept of truth completely in the eye of the beholder. In the following paragraphs, I will briefly discuss prominent theories of truth. I will not subscribe to a particular theory, nor is it necessary for the knowledge acquisition framework to strictly adhere to a particular theory. Discussed are the following theories:

1. Correspondence

2. Epistemic
   - Coherence
   - Consensus

3. Experientialist

3.2.1.1 Correspondence Theory

It is more correct to speak of the correspondence theories of truth, rather than a single theory, because this flavor of truth theory has seen many variations over time. In general, though, according to correspondence theories of truth, a statement is true when it corresponds to an actual state of affairs.
The correspondence theory is the most widely accepted amongst the theories of truth. Apart from its intuitive nature, it seems to guarantee that only those statements are seen as true that are reflected in a state of affairs.

An obvious objection to this theory is that the world and its actual state of affairs may not be completely accessible to us. Even from a realist’s point of view one will have to admit that large parts of the world are hidden from the observer. Thus, correspondence works well as a normative approach, at least in the positive case, because we can say with certainty that if a statement corresponds to the state of affairs in the real world, it is true. See for example Lakoff and Johnson (1999, p. 105) for a rebuttal of the correspondence theory.

3.2.1.2 Epistemic Theories

Epistemic theories of truth tend to conflate the issues of truth and justification of a statement. Normatively speaking, the two are not compatible, because a statement may be false despite there being sufficient evidence for its correctness on the one hand and statements can be true in the absence of any evidence for them on the other. The often invoked parable of the rooster that thinks it makes the sun rise because it crows every morning just before sunrise shows that even though the rooster is justified in believing that he makes the sun rise, this belief is simply false. The statement “there exists intelligent life outside the solar system” may be true or false, but its correctness is independent from our evidence for it.

However, as observers who do not have complete access to the world, methods of justification such as the scientific method are powerful tools to determine the correctness of a statement or a belief. Two ways of epistemic justification are those of coherence and correspondence:

**Coherence Theory** As the name suggests, the coherence theory sees the truth of a statement as coherence of the statement with a system of propositions. As with the Correspond-
dence theories, it is best to talk about Coherence Theories, rather than a single theory. The general idea, however, is that truth is a property of an entire system of propositions. The truth of an individual statement is then a function of its coherence with the overall system. The different flavors of this theory mainly disagree in the question as to whether there are multiple, possibly contradictory systems of propositions or a single, non-contradictory system.

Coherence is an essential criterion for truth in formal systems, such as logics and mathematics. Given true premises and a sound inference mechanism, all derived statements will also be true. The applicability here is however limited to closed systems. Truths that are merely derived from existing axioms were already implicit in the system and do not constitute new knowledge. Additionally, the absence of a contradiction is only a necessary, but not a sufficient condition for a statement to be true. Moreover, a set of incorrect assertions can be perfectly coherent, given that the premises or axioms we accept as true, are already false. Thus it is necessary to have a complementary justification procedure outside of the system that validates the premises.

Nevertheless, coherence is a valuable condition for computational systems as it allows to not only conclude a truth value for new statements, but also makes it possible to question premises when an assertion that has been accepted as true by some other means than by coherence, is considered to be false by the logic-based system. In that case, either the first validation mechanism or some of the premises are false and need to be revisited.

Consensus Theories Notably advocated by Jürgen Habermas (2003), consensus theories of truth are generally based on the classic idea of the consensus gentium (Latin for “agreement of the people”). As Vergilius Ferm states: “that which is universal among men carries the weight of truth”. However, a pure form of consensus is generally seen as insufficient to constitute an adequate theory of truth. Factors such as power structures, peer pressure or even superstition can cause a statement to be agreed upon by a majority of the mem-
bers of a group of people, even if the statement is factually wrong (and can even easily be proven wrong). If the goal is a universal consensus amongst all people, instead of a consensus within a group, then an agreement will be extremely hard to reach, even under ideal circumstances. Another criticism is that the consensus theory renders itself untrue by the simple fact that not everybody agrees on it.

However, as a descriptive theory, consensus has been fairly successful, as it can explain how statements are accepted as being true within a community. The consensus theory is thus often of sociological interest. Also, scientific communities tend to be more willing to engage in rigorous scrutiny of the presented facts and even their own premises. In practice, a consensus approach can therefore be successful, if the community is chosen carefully.

3.2.1.3 Experientialist Theory

An account of truth that is inspired by recent developments in linguistics and cognitive science is given by Lakoff and Johnson (1980, 1999). Their experientialist account of truth incorporates elements of correspondence theory, coherence theory, pragmatic theory and classical realism. Thereby we understand a statement as being true in a given situation when our understanding of the statement fits our understanding of the situation closely enough for our purposes.

Embodied truth: A person takes a sentence as “true” of a situation if what he or she understands the sentence as expressing accords with what he or she understands the situation to be. (Lakoff and Johnson (1999, p.510))

Lakoff and Johnson thereby claim that whereas embodies truth is not absolute objective truth, is not a purely individualistic account of truth, either, because the embodied experiences are shared amongst people.
Embodied truth is not, of course, absolute objective truth. It accords with how people use the word true, namely, relative to understanding. Embodied truth is also not purely subjective truth. Embodiment keeps it from being purely subjective. Because we all have pretty much the same embodied basic-level and spatial-relations concepts, there will be an enormous range of shared “truths,” as in such clear cases as when the cat is or isn’t on the mat.

(Lakoff and Johnson (1999, p.103))

The experientialist theory is of particular interest in the context of the creation of domain knowledge models, because these models usually reflect an individual point of view or the shared point of view of a group of people, for example a scientific community. Different from consensus theories, however, the notion of truth is not purely attributed to a rational agreement, but it is due to the shared and embodied experiences of the group members. An example of a knowledge model created by a scientific community that would fail according to a coherence theory, but is agreed upon by the community and true according to consensus or experientialist theories is presented in Chapter 4.

3.2.1.4 Conclusion

When truth is seen as an absolute, something that is independent from our inquiry into it, the correspondence theory is sensible. When we talk about knowledge, subjectivity comes into play with justification and belief (in the case of individual knowledge). This requires us to be careful with the way we talk about truth and knowledge. Often, we would have to say “I am justified in believing that X” instead of “I know that X” as well as “there is strong evidence for X”, instead of “X is true”. The more descriptive truth theories come closer to our linguistic understanding of truth, which is usually more of the form “I strongly believe that X”.

We may say that a statement is necessarily true, if it can be shown to correspond to a state of affairs. As such, truth by correspondence is a state that is desirable to achieve.
However, if correspondence cannot be established beyond the shadow of a doubt, it may suffice to either have a looser definition of correspondence that states that we should be sufficiently well justified in claiming the correspondence with the world or to use a truth condition given by another theory.

Each of the theories that were outlined here have strengths and weaknesses. The correspondence theory is a normative and absolutist theory of truth, whereas others are descriptive and relative to a specific context. Descriptive theories cannot guarantee the truth of a statement and we cannot always establish correspondence between a statement and a fact when applying correspondence theory. Thus, theories of truth are highly contested in epistemology. However, in this discussion the distinction between normative and descriptive often does not become apparent. Instead of seeing these theories as incompatible, maybe the epistemological discussion should be about how theories are of practical use. Correspondence should be a goal, an incentive to strive for further inquiry. When correspondence cannot be established, we need to opt for a descriptive theory that still guarantees a high certainty of correctness.

To give an illustration for the difficulty of assigning correspondence even for scientifically well-established facts, Chapter 4 will describe the development of an ontology that formalizes a domain as it is understood by experts. It is thereby acknowledged, however, that much of the understanding of the domain is actually an abstractions of reality. An ontology or domain model that represents the human conceptualization of a domain is thus itself a representation of a metaphor, rather than give a representation that truly corresponds to reality. In such cases, the experientialist theory allows to call the ontology a true representation.

Regardless of the theory that is espoused, the general consensus is that statements that are untrue should eventually be identified as such. Thus it is important that statements can always be scrutinized. The hermeneutic method of knowledge acquisition in this work lays the conceptual foundations that allows beliefs to be revisited and revised, if necessary.
3.2.2 Justification

To be called knowledge, a belief does not only have to be true, the believer should also have a good reason for believing it, i.e. a justification. Justifying a belief can take various forms. The belief can e.g. be

1. self-evident
2. coherent with an already held set of beliefs
3. confirmed by external sources
4. created by a trusted procedure

In this work I will show how criteria (2)-(4) are used within the system. The validity of these criteria can be seen as parallel to some of the truth theories described earlier. Criterion (2), for example, justifies a statement, if it logically fits within the set of already held beliefs and thus justifies a statement if the coherence theory of truth is accepted as valid. Criterion (3) is analogous to a consensus approach to truth. However, the external sources may have established correspondence between the assertion and a state of affairs and feel thus justified in validating the statement. Criterion (4) is often shown inductively. If a procedure has been successful in creating true statements in the past, it is likely that it will be successful in the future. Here, this criterion mostly applies to automated information extraction techniques. It is important to note, though, that a single type of justification is not seen as sufficient. For example, the trusted information extraction procedure needs to produce a statement with high certainty and the statement needs to be coherent with the knowledge base and/or it needs to be confirmed by external sources.

Justification, unlike truth (when seen from a realist point of view), is a cognitive and a social or systemic process. In the case of a coherence approach, the justification may be externalized to an algorithm, but that is the exception. Lipkin (1992, p.607) defines justification as “consensus by ‘qualified’ individuals.” This enables justification within a
manageable set of peers, that is, a statement should be accepted beyond reasonable doubt, but not necessarily by every member of a group.

In the *Wisdom of the crowds* paradigm (Surowiecki, 2005), the definition of “qualified” has been extended beyond people we would usually consider to be experts. Depending on the problem, any sufficiently large group of independent individuals can render qualified judgements.

### 3.2.3 Belief

A very good brief introduction to the topic of *Belief* is given in (Schwitzgebel, 2010). Here, I will only touch the definition of belief with respect to its important to the question of knowledge acquisition in a system. Traditionally, a belief is seen as a state of mind that is mostly described as a *Propositional Attitude*. It is thus intuitively something that an individual can have, but not a group or a system. However, more recently, a distinction has been made between *subjective* and *objective* beliefs Bird (2010), i.e. beliefs that individuals hold internally and beliefs that are objectified, for example by writing them down in text or a formal language. Thereby, systems are naturally able to hold beliefs given a belief is a formal statement that the system can access and reason about. This view is a necessary precondition for this discussion and is covered in-depth in the next subsections.

### 3.2.4 Knowledge in a Group - Social Epistemology

The categorical definition of knowledge given thus far is traditionally aimed at a subjective account of knowledge held by an individual. Social epistemology focuses on the knowledge held by a group. Here, a distinction can largely be made between two camps:

1. Knowledge held by a group is the accumulation of the knowledge held by the individuals in the group
2. The group can hold knowledge that is distinct from the knowledge held by the individuals

According to Bird (2010), social knowledge is seen as functionally analogous to, rather than dependent on individual knowledge. A group thus becomes a knowledge-bearing agent in its own right. In my opinion, however, Bird makes a stronger case for a Systems Epistemology than for a group Epistemology. In his view, a group becomes a knowledge bearer once the knowledge is accessible to the group in the form of e.g. a publication (I will get back to this argument in Section 3.2.5). This means that social knowledge requires the physical state of knowledge accessibility rather than a more intangible state of mind(s). I believe a stronger argument is touched in the paper, but not pursued properly. Bird talks about how the group of scientists at the Manhattan project worked together to develop the first atomic bomb. Due to the shared nature of the work, no single scientist had the knowledge to build a bomb, but the group as a whole had that knowledge. Whereas in the case of externalized knowledge (scientific papers, formal representations of facts, etc.) the group may or may not have a particular “state-of-mind” of knowing, the Manhattan project example shows that this state of mind can exist for a group. Every scientist in the group knew that they had the knowledge to build the bomb and every scientist also knew that the combined knowledge of the group was needed.

An interesting interjection to the discussion of whether knowledge can be held by a group as an entity or rather by their individual members is given by Surowiecki (2005). Even though he mostly talks about problem solving, I think the hypothesis can be expanded to beliefs or knowledge. In many of the cases he describes, very few individual members of a group were able to solve a problem optimally and no member of the group was able to consistently outperform other members. However, the aggregated answers always gave a close-to-optimal answer to the problem. This is an indication that, given the right aggregation framework, a group could be said to hold a belief that is different from the beliefs of every member of the group.
3.2.5 Knowledge in a System - Systems Epistemology

The term *system* stems from the Latin *systema* that derives from Greek σύστημα, which literally translates to *composition* and is used in the sense of a *whole compounded of several parts or members*\(^1\). The World Wide Web is a massive system that is comprised of billions of machines, billions of users and an unbounded number of artificial agents that all perform tasks on the data and information available. In this work, the view of a system is that of interaction between human and artificial agents, specifically with respect to knowledge acquisition.

The idea of machines processing information similar to a human agent has been famously brought forth by Newell and Simon (1976), but the extent of the comparison has since been met with criticism. The goal shifted from creating machines that think like humans to a) machines that assist humans or take on small tasks in which they outperform humans and b) machines that collaborate with humans in problem solving (Von Ahn and Reiter, 2005; Thomas and Sheth, 2011; Wegner, 1997). The *Semantic Web* (Berners-Lee et al., 2001), even though initially mostly thought of as a platform for machine-machine interaction, aims at providing a general framework for this systemic interaction that involves humans and machines.

Both views held in social epistemology - a) group knowledge as a function of the individuals’ knowledge and b) group as a knowledge bearing agent, can be used as starting points for a systems epistemology. Instead of a dichotomy, we should maybe think about a systemic interdependence, in which individuals affect the group’s “state of mind” and in which the individual agent is affected by the group. Hence, instead of asking *who* can bear knowledge in a system, we need to ask *how* a system can bear knowledge. In order to speak of a systems epistemology, we need to accept the notion of objective knowledge that can exist independently from a subjective knowledge bearer, for example in books or

\[^1\]http://www.perseus.tufts.edu/hopper/morph? l=susthma&la=greek &can=susthma0#lexicon
as formal statements in an ontology or a knowledge-based system. Objective knowledge mostly corresponds to the subjective knowledge that can be made explicit as statements or predicates represented in a machine-readable way.

There are two basic approaches to define the concept of “knowledge”, knowledge as a thought in the individual’s (or subject’s) mind, and knowledge as an object or a thing. The first approach conditions the knowledge in the individual’s mind. Knowledge is a thought. It is characterized as “a justified true belief”. This definition of knowledge as a justified true belief is originated from Plato’s Theaetetus (Plato and Campbell, 1883) [...] it seems sufficient for our purposes to characterize subjective propositional knowledge by the certainty of the individual that his/her own thoughts are true, and his/her ability to base this certainty on a sound justification. Note that in the subjective domain “knowledge” is the content of a justified true thought in the individual’s mind, while “knowing” is the state of mind, which is characterized by the three conditions: justification, belief, and truth.

The second approach ascribes an independent objective existence to knowledge. Knowledge is the meaning, which is represented by expressed propositions. It is true and exists independently of, not depending on, subjective knowledge of the individual knower. The implications of this approach to [Library Information Systems] were recently discussed by Hjø rland (2004).

(Zins (2006))

Zins’ second consideration is important for a systems epistemology, even though it is not clear to me why these two approaches are juxtaposed, rather than synthesized. As long as belief does not have to be characterized as being a state of an individual’s mind, knowledge does not have to be held by a single individual Bird (2010), but can be a formal representation of justified and true statements that are held or managed by a system. Thus,
in a system that aims at acquiring knowledge, it is important to have components that can hold beliefs, that can assess the truth of a statement and that deploy methods to justify it.

Nonaka and von Krogh (2009) responds to criticism from traditional views that knowledge needs to be subjectively held:

[...] we think that the term ‘knowledge’ should apply if it results from the justification of belief and if it enhances the capacity to act, define, and solve problems [...]. At one extreme of the continuum, some simple explicit knowledge can even enable machines to solve very specific, constrained, and well-defined problems. As Dreyfus et al. (2000) convincingly argue, expert knowledge can never be fully captured in computer software due to the tacit and embodied elements. Yet, expert knowledge is a basis for increasingly explicit knowledge on which to create automated processes.”

(Nonaka and von Krogh (2009))

Nonaka and von Krogh (2009) invoke the issues that come with the presence of tacit knowledge in a system. However, they also realize the potential of explicit (objective) knowledge in further automation. In this dissertation I will show that whereas it is extremely difficult to automatically extract tacit knowledge, it is possible to formalize it (to a large extent) in a social process and then further use this formalized tacit knowledge to improve automatic extraction and enrich the automatically extracted knowledge with elements of the tacit knowledge. Chapter 4 presents opportunities and challenges that come with the formalization of tacit knowledge.

Once it is accepted that knowledge can be held by a group or by a system, it is important to note that a group or a system can have different degrees and qualities of knowledge. Halpern and Moses (1990) argues for the distinction of 5 types of knowledge within a system or a group. Here these five are seen as referring only to objective (explicit) knowledge. The term “objective” is used here rather than the term “explicit”, because Halpern uses the term “implicit knowledge” to mean “knowledge that can be inferred through deductive
1. $I_{GP}$ - The group $G$ has implicit knowledge of $p$ (not to be confused with tacit knowledge), i.e. if one part of the system knows that $p \rightarrow q$ and another part of the system knows $p$, then the system implicitly knows $q$

2. $S_{GP}$ - Someone in $G$ knows $p$.

3. $E_{GP}$ - Everyone in $G$ knows $p$

4. $E_{GP}^{k}, k > 2 - p$ is $E^k$-knowledge in $G$

5. $C_{GP}$ - $p$ is common knowledge in $G$

$S_{GP}$ and $E_{GP}$ are self explanatory. $E_{GP}^{k}$ means that not only does everybody in the group know that $p$, but “everyone in G knows that everyone in G knows that ... that everyone in G knows that $p$ is true” is true, where tile phrase ”everyone in G knows that” appears in the the sentence k times.” (Halpern and Moses, 1990, Page 2). This means that the group or the system is to a degree of $k$ aware of its common knowledge. $C_{GP}$ means that $E_{GP}^{k}$ holds for all $k$, i.e. the group or system has a full awareness of all its members/components knowledge regarding $p$.

The idea of implicit knowledge is particularly intriguing, because it concerns knowledge that is not yet out in the open, but could easily be discovered if the different components or agents that are part of the system were to interact and share their knowledge or their beliefs. Once these beliefs or pieces of knowledge are made explicit, i.e. are formalized, this is an easy task in an Information System. In the Semantic Web Vision, the interaction of agents based on formalized knowledge is a centerpiece (Berners-Lee et al., 2001). This vision has been impeded to some degree by the lack of formalized knowledge, but mostly because the available knowledge could not easily be understood by all participating agents. Different formalizations of the same fact are not easily mappable, so even though $S_{G}(p \rightarrow q)$ and $S_{GP}'$, the system may not know that $p$ and $p'$ are the same
propositions. Proper mapping of statements in the absence of complete knowledge is much easier for humans than for machines, therefore in a collaborative system it is more likely that \( S_G(p = p') \) and thus the system can derive \( q \).

The "Circle of Knowledge" idea that is presented here gives a possible solution to this problem. Independent agents add pieces of knowledge whenever they are interacting with the system, which can include mapping two concepts or statements that were previously not known to express the same concept or fact. Additionally, the mapping problem is often caused by reference problems, which are addressed in the next section. In the context of automatic knowledge acquisition, reference problems should be avoided at acquisition time.

### 3.3 Reference

One of the problems that automated methods to Information Extraction from text are facing is that of reference. Welty and Murdock (2006) treat it as a problem of entities vs. annotations. I prefer the terminology concept vs. concept mention. The issue here is first whether a term in text refers to a concept, and secondly to which concept it refers. It brings about problems of disambiguation and co-reference resolution. A term in text is not the concept or the entity itself, it merely is a symbol that can refer to it. Henri Magritte made this vividly obvious in his painting of a pipe. The painting shows a pipe, but the writing on the painting says “Ceci nest pas one pipe” (This is not a pipe). At first confusing, it soon becomes apparent that the painting is a symbol for a pipe, i.e. merely depicts a pipe, rather than being a pipe. In chapter 5.3 I will show how in the domain definition, i.e. the concept base, the reference problem can be alleviated by extracting concepts and entities from a conceptual corpus. Having already identified pertinent concepts, it is much easier to identify their mentions in free text, rather than inferring the concepts from text. Assuring proper reference will also help avert Gettier-type problems that were mentioned in Section 3.2.
This work follows a view that concepts and entities can be identified both by rigid
designators (Kripke, 1980) and by definite descriptions (Russell, 1905). This view is un-
derlying the 2-step creation of domain models encompassing domain definition and domain
description. In the following, I will briefly outline the ideas behind rigid designators and
definite descriptions. A more in-depth discussion with respect to naming concepts on the
Web can be found in Halpin (2010).

3.3.1 Rigid Designators

A rigid designator of a concept refers to the same concept in every possible world. This
means that there is a necessary correspondence between the name (i.e. the designator) and
the concept. The semantic content of the designator is thereby only the reference to the
actual concept or entity, it usually does not have any descriptive content.

For example, the name Saul Kripke will refer to the same person in every possible
world in which Saul Kripke was born. The description “Saul Kripke, a famous philosopher
of language” only describes an incidental property that does not hold in a world where Saul
Kripke was e.g. an aerospace engineer or a janitor.

3.3.2 Definite Descriptions

A definite description is a phrase that properly and unambiguously denotes an entity or a
concept by virtue of a property that the concept holds. It is usually of the form “The F is
G”, such that $\exists x (F(x) \& \forall y (F(y) \rightarrow x = y) \& G(x))$. This can be broken down as follows:

1. There is an $F$ (Existence)

2. At most one thing is $F$ (Uniqueness)

3. Something that is $F$ is $G$ (Universality)
For example the phrase “The first man on the moon” is a definite description for Neil Armstrong, “The 44th president of the United States of America” is a definite description for Barack Obama, and so on.

There have been a number of criticisms of Russell’s theory of definite descriptions that can be broadly divided into two approaches. The first is challenging the truth conditions of Russell-style descriptions, especially in cases when the entity that is described does not actually exist or when the descriptive property is not a necessary property of the entity. The second one is challenging the notion that a description can ever be complete. According to Kripke (1977), this second criticism of incompleteness is enough to discredit the theory of definite descriptions as not sufficient to unambiguously identify concepts and entities. However, recent interpretations of Russell’s theory suggest that his theory is compatible with Kripke’s De Castro (2007),

Apart from not assuring that definite descriptions unambiguously identify a concept, the theory is very restrictive when it comes to reference. Many statements that we as readers or listeners understand perfectly well are in fact non-referential according to Russell. Often we understand them metaphorically or as hyperboles. For example Phillip J. Fry’s (Futurama) “Nobody drove in New York. There was too much traffic.” This sentence could, however, not be literally translated into a formal statement and still make sense. For this reason, the theory of definite descriptions provides a fitting theoretical foundation for Domain Description.

3.3.3 Application

Accepting the definition of an ontology as a “specification of a conceptualization” (Gruber, 1993a), it follows that the concepts that are formalized in the ontology are representations of a conceptualization of real-world concepts. This means that there is a fiat correspondence between the concepts in the ontology and the real-world concepts or entities. The names of concepts used on the Semantic Web are URIs (Halpin, 2010). Wikipedia can be seen as a
conceptual corpus where concepts and entities are identified and rigidly named in the form of URIs. An automated extraction of concept identifiers from such a corpus again yields rigid designators for the extracted concepts.

The strategy of splitting the extraction of formal models into the two parts of (1) extracting concept references from an existing conceptualization in the definition part and (2) extracting a concept description of a domain allows the two theories of reference to be deployed together. The theory of rigid designators states that a name necessarily references a concept in every possible world the concept is present in. Since it only constitutes a reference, though, it is devoid of attributes and relationships, which constitute descriptions of the concept in the current world. This is beneficial for a rigid concept reference, because many of those would be incidental properties. However, our knowledge about concepts usually comes from descriptions.

The mere knowledge of the existence of a concept is not helpful when we do not know of its attributes and relationships with other concepts. The combination of both is essential for a useful knowledge representation. The modality does thereby not matter, because we are interested in the knowledge of the current state of affairs. Moreover, providing a conceptual basis with the Domain Description assures that no descriptions of non-existents are extracted. Otherwise the above example could have been formalized as \( \langle \text{Nobody} \rightarrow \text{drives in} \rightarrow \text{New York} \rangle \).

Since both Russell and Kripke thought of their theories as normative, neither spoke of the synthesis of their theories. However, in practice, definite descriptions can help disambiguate when concept mentions are not rigid but rather ambiguous. (Yu et al., 2011) address this kind of mutual disambiguation in an IE scenario.

For example, the facts that Neil Armstrong was an astronaut and the first person to set foot on the moon will be of great importance for the knowledge base. To properly handle those facts, however, it is important that the knowledge base entity “Neil Armstrong” correctly identifies the person who went to the moon, rather than e.g. the Canadian Ice Hockey
player “Neil P. Armstrong”.

One practical challenge is to bring the two different extraction techniques together in a meaningful way. When all that is available as a concept identifier is a URI, then only concept descriptions that also contain this URI can be used to enrich the concept. Since I am interested in expanding the amount of available knowledge about the concepts in question, just taking descriptions from a corpus that already contains formal information about concepts (e.g. in RDF format, as in DBPedia), is not enough. New information needs to be extracted from text. The URI designation is naturally not present in free text, so different designators need to be found. Chapter 5 will describe the technical details of this approach. Suffice to say here that terms need to be found that identify a concept. Context is thereby of importance as many terms refer to more than one concept. Hence a measure that indicates a degree of ambiguity needs to accompany a term. In the implementation, a probabilistic approach was chosen to reflect the uncertainty of reference.

3.4 The Hermeneutic Circle

In the introduction chapter I mentioned the idea that my approach to knowledge acquisition in a system is following the metaphor of the *Hermeneutic Circle*. The idea of hermeneutics was conceived in ancient greek philosophy as a way to interpret linguistic and non-linguistic expressions (Ramberg and Gjesdal, 2009), but has changed over time to become a universal theory of interpretation that covers theological, linguistic, philosophical and other areas. Recently it has been applied to Information Systems Epistemology Hirschheim (1985). Regardless of the application field, hermeneutics is based on the idea that an understanding of a *whole* is dependent on the understanding of its individual *parts*, which, in turn, can be better understood in the context of the *whole*.

As a metaphor for continuous knowledge acquisition in a system, the Hermeneutic Circle expresses the idea that the pursuit of knowledge is an ongoing endeavor. Given the
knowledge that a system holds at any given point in time, it can apply this knowledge to interpret new information. Moreover, any fact that the system already “knows” is subject to constant verification as more knowledge becomes available.

3.5 Knowledge Acquisition in a system

Having established the epistemological foundations for acquiring knowledge individually, in a group and in a system, I will outline a conceptual approach to achieving this goal in a large system such as the World Wide Web. In the previous definition of knowledge it was shown that the most contentious element is that of truth. Epistemic approaches mostly reduce the act of truth finding to that of strong justification, in which case knowledge merely becomes “justified belief”, which will be unacceptable to many advocates of correspondence.

In the absence of a clear function of correspondence in many cases, however, and in the light of a system that can hold and revisit beliefs, it may be acceptable to define knowledge for the purpose of this work as “Continuously justified belief”. This means, an assertion in the system is valid, if, whenever it is scrutinized, there is ample evidence for its truth and it is accepted by a group of peers who are qualified to make this assessment.

The cyclic nature of knowledge acquisition has been discussed extensively in the literature (Popper, 1963; Nonaka, 1994; Tress et al., 2006; Velasco et al., 2011). The circle of knowledge (Figure 3.1) that was briefly mentioned in the introduction is an abstraction of several theories of knowledge acquisition. Some theories include a knowledge sharing stage (Velasco et al., 2011). In the work described here, sharing is an integral part of the validation and the storage stage. A communal verification requires sharing of and interaction with the knowledge in a social setting. The verified knowledge is also assumed to be publicly available and is thus automatically shared.

The circle can be seen as representing a cyclical version of Popper’s (1963) theory of
the progress of scientific knowledge, in which initial conjectures are meant to be rigorously scrutinized and those that withstand the scrutiny form the current state of knowledge. That does not mean that the scrutinized statements or conjectures mark the final state of knowledge. Every statement can always be subject to verification. Here, the rigorous scientific inquiry is replaced by the general validation step that is theoretically open to any kind of validation and is practically dependent on the field from which statements are validated, the particular implementation of the knowledge acquisition task and the expertise of the individuals involved.

In the descriptions of the different theories of truth it became obvious that each of them can either allow assigning truth to an incorrect statement or, in the case of correspondence, may not be able to allow the assignment of a truth value yet. In the hermeneutic view of knowledge acquisition the constant verification and revisiting of already asserted statements is an intrinsic part of the system, which assures that the truth values of statements conform to the latest state of affairs. The collaborative character may favor a consensus
approach to truth, but it can always be challenged by a coherence requirement.

The circle that is presented here does not make a metaphysical claim about the character of knowledge. The notions of present knowledge, learning/information acquisition and validation are applicable to many theories of knowledge. A shortcoming of this abstraction is that it does not account for epistemological and ontological dimensions that qualify the kind of knowledge that is acquired. In particular, I do not consider tacit knowledge in the circle, because I assume a system that is loosely coupled and where the communication between the stages of knowledge acquisition happens only by virtue of formal statements. Nonaka (1994) adds these dimensions in his theory of organizational knowledge acquisition (see Figure 3.2). In his view, as knowledge acquisition progresses through different stages of an organization, there is an interplay between explicit and tacit knowledge until sharing of knowledge is required in an inter-organizational setting.

Much of the knowledge that deserves a formal representation is expert-knowledge. This is to a large extent, because general-purpose knowledge is already available in the form of Linked Data or in CYC. Expert-knowledge does not necessarily mean scientific knowledge, but knowledge that is held by a limited number of people, but can become important for a large number. This can include scientific facts, but also facts about how to fix a broken faucet.

To get “into the heads” of experts to get a formal account of their knowledge, we have two choices. The first is to work with the experts directly and “extract” their knowledge in a knowledge engineering fashion. Chapter 4 shows one possible implementation of a knowledge engineering approach to knowledge acquisition. A knowledge engineering approach allows the engineers and the experts to also formalize tacit knowledge (and thus make it explicit), often just by requiring a complete chain of reasoning behind a stated fact. The second choice is to use secondary knowledge, i.e. the knowledge that has been written down by the experts, and use manual or automated ways to formalize the knowledge contained in text. This is particularly important because experts may be too busy to engage
3.5.1 Practical Considerations

Even though most of this dissertation will describe the theory and practice behind the implementation of a knowledge-acquisition system, it was important to lay out an epistemological background for knowledge acquisition within a system. On the one hand it is necessary to be aware of the fact that information is not the same as knowledge. But more importantly, it makes a clear case for the necessity of having constant verification of the statements that a knowledge-based system operates upon.

This chapter thus laid a background for a methodology that makes it possible to call the acquired statements knowledge, given that the requirements of belief, justification and
truth are met. However, these requirements need to be attainable within the system. Hence, I make the following assumptions for a working system:

- **Belief** - The system holds knowledge in the form of formal statements.

- **Justification** - Validation measures are accomplished in a manner that allows justification. Thus, many people have to agree on a statement for it to be considered in the first place. Further, the algorithms used for information extraction need have a high degree of precision. This provides a high degree of confidence in the correctness of extracted statements.

- **Truth** - It is assumed that the information sources used in this work contain mostly correct information. It thus needs to be shown that the aggregation mechanisms that are used are truth-promoting. There is also a - somewhat overoptimistic - notion that one of the first principles of human nature is “a propensity to speak the truth” (Reid, 1764), which would indicate that humans tend to speak the truth more often than not. However, even if this notion is only statistically correct, an aggregation of statements will likely yield a correct outcome. This idea is also underlying the “Wisdom of the crowds” (Surowiecki, 2005) paradigm, which gives good indications that a statement that has been asserted by many independent agents and/or been validated by independent agents is likely to be correct.
Knowledge Engineering - Based Domain Model Creation

The world is the totality of facts, not of things. (Wittgenstein (1922))

This chapter describes a knowledge engineering approach to knowledge acquisition in a tightly coupled system. It provides a contrasting view to the loosely coupled knowledge acquisition system that is presented in chapter 5. Here I demonstrate how a domain definition and part of the domain description can be created manually based on expert agreements about the knowledge in a domain. However, an automated domain description algorithm is also used to add structural and factual knowledge.

This chapter is meant to contrast knowledge acquisition for explicit knowledge with an attempt to encode a combination of tacit and explicit knowledge that constitute a deep understanding of the domain at hand. Many of the triples that encode knowledge in the ontology that is described here could not be extracted using the general-purpose methods described in chapters 5.3 and 5.4.

By the time the research that underlies the chapter was conducted, the Web Ontology
Language OWL was still in its first iteration. The next iteration, OWL2, introduced some new features, such as punning, that could potentially change some of the formalizations in the ontology.

4.1 Introduction

The field of BioInformatics has seen a dramatic increase of available ontologies for many of the life sciences domains. The Ontologies in the OBO project\(^1\), especially the Gene Ontology (GO)(Ashburner et al., 2000) with its comprehensive schema and thousands of instances, take leading roles. As a broad lexicon or dictionary, GO serves one of the major purposes of ontologies: facilitating agreement. However, it is not designed for extensive computational use, so the amount of inference that can be done with the knowledge is limited. Only two types of relationships between the entities in the ontology are formalized: $is\_a$ and $part\_of$.

An ontology that provides rich, machine accessible relationships must be rigidly formalized. Knowledge modeling languages such as KIF (Genesereth et al., 1992), RDF (Klyne et al., 2004) or the W3C-recommended Ontology Web Language OWL (Horrocks et al., 2003) allow such formalizations with different expressiveness. OWL promises to be a good compromise between expressiveness and computational complexity on the one hand and versatility and simplicity on the other.

This chapter focuses on issues related to representation, expressiveness, granularity and instance population in the development of the Glycan Structure Ontology GlycO. It is one of the ontologies designed as part of a suite of web-accessible ontologies for the glycoproteomics domain alongside the Enzyme function ontology EnzyO and the provenance ontology Propreo (Sahoo et al., 2006). The goal of this suite is to have a basis for description, annotation and reasoning, such that every step from experimental setup

\(^1\)OBO: Open Biomedical Ontologies
over experimental conduct and analysis to acquisition of hypotheses and theories can be formalized. This work was conducted in the context of the “BioInformatics for Glycan Expression” core of the NCRR Integrated Resource for Biomedical Glycomics project at the Complex Carbohydrate Research Center (CCRC) of the University of Georgia.

Glycans are complex carbohydrate structures, which play key roles in the development and maintenance of living cells. Glycans are built from simpler monosaccharide residues (such as mannose and glucose), which constitute the nodes of tree structures with edges that are comprised of chemical bonds between the residues. The synthesis of these glycans in organisms is an intricate process that can be modeled as a collection of biosynthetic pathways. At each step in such a pathway, an enzyme-catalyzed reaction “adds” a new residue as a leaf to an existing structure or “moves” a whole sub-tree to a different parent. It is well established that alongside genes and proteins, glycans play a major role in cell functions.

The aim of glycoproteomics is to understand cellular processes that are mediated by the interaction of proteins, the genes that encode them, and the glycans that are attached to them. The goal in developing GlycO has been to assess the extent to which knowledge in this domain can be logically formalized to facilitate the discovery and specification of relationships between the glycan structures, their metabolism, and their functions. Among the challenges faced were those of a limited expressiveness of the chosen OWL-DL standard, and mereological issues of granularity.

The main contributions of this work include:

- Creating a more meaningful domain model by
  - Building a Domain Definition that captures the richness of the domain using expressive language, esp. restrictions
  - Supporting modeling of molecular structures that are important for domain scientists
– Rigorous modeling with contextual archetypal instances used as building blocks

- Creating a Domain Description for the ontology by extracting and disambiguating instance information from multiple heterogeneous sources

- Allowing for more meaningful queries by formalizing knowledge that is usually inferred in database models

- Addressing granularity issues

Following this introduction, section 4.2 will describe the conceptualization and formalization of the glycoproteomics domain in GlycO. section 4.3 will detail the sources and algorithms used for the automatic population; section 4.4 will evaluate GlycO and discuss the impact it can have on biochemical applications. Section 4.5 finally concludes the chapter.

### 4.2 Ontology Design

#### 4.2.1 General Considerations

The rules of syntax alone cannot determine the meaning of the statements expressed by the words in that syntax. A fundamental aspect of ontology development is the capture of semantics in a formal syntax, i.e., the unambiguous formalization of statements or states of affairs. Representation of meaning using first order logic is limited to stating that an object has certain properties and relationships with other objects. Even generalizing these properties to sets or classes of objects bears problems (Smith et al., 2005). It is necessary to find a balance between the unambiguous representation of objects including their relationships and any attempt to capture the infinitude of relationships present in the world.

We therefore are limited to modeling very specific problems that require a finite amount of representation. The critical objects and their relationships must be identified
and then formalized so that machines can infer new or implicit knowledge from the given information.

Collections of biological entities, such as genes, proteins and carbohydrates, are assumed to have a syntactic structure, much like natural language. For example, we assume that the structure of the genome directly or indirectly encodes the structure of the entire organism. By knowing the syntactic and semantic rules that govern gene structure, we can assign meanings to DNA strings and substrings, i.e., identify genes and the protein sequences they encode. Of course, this is not always a trivial task, but provided the genes themselves (and not their environmental context) constitute the information basis, we can gain a large amount of knowledge by studying gene syntax. A similar simplifying assumption is made for glycans, which clearly influence cellular properties: The correspondence between a glycoprotein’s biological properties and the presence of specific glycan structures at specific locations on the protein’s surface should be captured in the glycan’s formal description.

Developing a highly expressive formal ontology for a comparatively narrow field of research requires the constant interaction between domain experts and knowledge engineers. The modeling of knowledge calls for a profound understanding of a domain. The domain expert must fully participate in ontology development and understand the formalisms used for specifying the conceptualization of the domain. Conversely, the knowledge engineer must analyze the ontology to avoid ontological fallacies in modeling. As shown in (Johnson, 1983), it is important to make the domain experts’ tacit knowledge explicit. The experience in the GlycO project showed that this can only be done gradually, by having both the knowledge engineers and the domain experts come to a common understanding first. Then, as concepts are modeled, repeated questioning reveals the implicit assumptions that domain experts make about their field.

Finally, the knowledge engineers need to assure the formal correctness and consistency of the ontology. Of great help is the Ontoclean methodology (Guarino and Welty,
which explains how concepts should be classified on a meta-level according to distinctions like rigid versus non-rigid concepts, entities versus relationships, etc. At this step, the knowledge engineers need to scrutinize every concept together with the domain experts again, but from a formal perspective. The engineers need to explain classification rationales to the domain experts and see if the resulting modeling still meets the experts’ view of their field. These modeling steps may be repeated several times, until a consensus is reached.

Although GlycO is focused on the glycoproteomics domain, it is critical that it is sufficiently comprehensive to invoke important concepts in the related disciplines of proteomics and genomics. By providing links to other ontologies that describe the fields closely related to glycoproteomics, it allows for scientific discovery of complex or unknown relationships across research fields. Because it is assumed that the ontology will be used for such discovery, it needed to be strongly restricted to clearly distinguish the asserted concepts by semantically modeling the subtle differences in glycan structure that modulate their biological functions. Only then a correct identification of discovered concepts and relationships can be achieved. GlycO is meant to be more than a controlled vocabulary; its intention is to formalize the knowledge that is tacitly held by glycobiologists and to be used for reasoning in scientific analysis and discovery.

Making an ontological commitment to OWL 1-DL meant that a clear separation needed to be made between classes and instances. What seems like a straightforward ontological separation quickly becomes problematic when the different kinds of inferences that can be done on T-Box vs A-Box are taken into account. When a specific type of complex carbohydrate is described as a class, then it is only possible to make general assertions about the relationships that instances of this class can engage in. From a realist perspective, it is also not possible to model specific single instances of a molecule, because we cannot really establish a connection between the ontology and a real-world instance of a molecule.

\footnote{Punning, introduced in OWL 2, allows the use of an identifier both on the class level, as well as on the individual level. However, internally, the reasoner still clearly separates between the same ID used in the sense of a class versus used as an individual.}
especially not when we also want to model the molecule as part of chemical reactions.

Therefore, it is assumed in this work that the way scientists reason about chemical structures is by means of mental archetypes. For example, water is composed of two hydrogen and one oxygen atom. However, a reaction that creates water from hydrogen and oxygen is not necessarily thought of in terms of single participating real-world atoms and molecules. Rather, it is thought about as mental representations of single atoms and molecules that in turn represent two parts of hydrogen reacting with one part of oxygen. After giving an overview of the general considerations for the Domain Definition, Section 4.2.3 will more deeply discuss the concept of archetypal instances that are thought of as formal specifications of mental representations of the corresponding entities in the world.

### 4.2.2 Domain Definition - Schema Design

With the automatic extraction of chemical structures in mind, it was important to build a comprehensive and tightly restricted T-Box, so that an incorrect logical classification of extracted structures could be ruled out.

Initially, the glycoproteomics domain was broadly analyzed, terms were collected, and the way these terms are used by scientists was examined. It turns out that the informal use of the is_a relationship, as in “a glycan is a complex carbohydrate”, implies a hierarchy of concepts with multiple inheritances. It is also desirable to keep the “colloquial” use of the biochemistry terminology consistent with the ontology, while also adding more distinguishing descriptions in the form of named relationships and their restrictions. There are many ways of classifying monosaccharide residues, which are the building blocks of glycans. For example, it is possible (and equally valid) to classify them according to the number of carbon atoms in the monosaccharide or as a structural variant. That is, a $\beta$-D-Glcp residue can be identified amongst other criteria both as a hexosyl residue (with 6 carbons) and as an aldosyl residue (embodying the bothaldo structural variant). All of these properties are accounted for by allowing a particular monosaccharide residue to inherit from several super
classes. Whether this directed acyclic graph is explicitly asserted or subsequently inferred is secondary. For example, the absolute configuration D and subsumption by the super-class residue are necessary and sufficient properties of the class D-residue. A reasoner will automatically subsume any residue class that has the absolute configuration D under the class D-residue. A hierarchy with multiple inheritance will almost always automatically arise when a more sophisticated logical description of classes is used alongside restricting conditions. For this reason, criticism of multiple inheritance, as in (Soldatova and King, 2005) seems impractical.

The first level of abstraction contains the three classes “Chemical Entity”, “Chemical Property” and “Reaction”. This is an appropriate starting point in that upper level ontologies such as SUMO distinguish between “Object”, “Attribute” and “Process”. GO uses cellular_component, biological_process and molecular_function on the first level of abstraction. The analog to molecular_function is in our case defined in the functional ontology EnzyO, which describes enzymes and their functions. This compliance with standard classifications facilitates the integration of GlycO with other ontologies. From there, a finely grained class hierarchy is defined (see Figure 4.1 for a selection of the first 3 levels of the GlycO hierarchy).

The relationship hierarchy in GlycO is built with respect to emerging standards in the biomedical domain. The OBO relationship ontology (Smith et al., 2005) is used as a starting point and more refined named relationships are added. See Figure 4.2 for a part of the GlycO relationship hierarchy. With 14 levels, GlycO has a deeper hierarchy than many other domain ontologies. This finely grained class design is essential for the purposes of evaluating experimental results using the knowledge stored in the ontology. Small differences in the glycan structure might affect the kind of interactions an individual glycan or members of a class of glycans have with other objects in the ontology.

The hierarchy of concepts is one aspect of semantics captured in an ontology, but the
addition of other relationships is required to realize an expressive model. A concept by itself might be useful for a human observer, but only by understanding it within a context of other concepts. Scientists infer related concepts according to their background knowledge. For machines, this background knowledge needs to be stated explicitly. Soldatova and King (2005) raised the issue that the biomedical ontology MGED contained too many named relationships that impede the computational use of the ontology. I disagree with this assessment of ontology design. A large number of named relationship increases the semantic value of an ontology (Sheth et al., 2004), if these relationships are well defined. The dilemma of generality versus computational complexity is instead addressed by making use of a relationship hierarchy, i.e. modeling the relationships from more general down to more specific. Upper level relationships are e.g. has_part or affects and their inverses. Inheriting lower level relationships restrict domains and ranges of the upper level relationships. For example, has_carbohydrate_residue is essentially a has_part relationship, but its domain is restricted to glycan and its range is restricted to carbohydrate_residue. If the ontology is to be merged or aligned, an alignment algorithm will be able to map this relationship to a more general relationship in a different ontology that does not explicitly formalize the specific has_carbohydrate_residue relationship.

As the name indicates, a class hierarchy provides a means of classification. Together with relationships and restrictions it specifies what can possibly exist within the realm that is described. Classes themselves exist only in a very abstract sense. The instances in the ontology are meant to provide a representation of the things that actually exist in the domain of interest.

### 4.2.3 Archetypal Instances

Ontologies have gained high visibility in the life sciences, especially in the biomedical field. The NCBO and OBO ontology portals grow rapidly and large efforts are made to
have a unified underlying ontological model. For applications within the Semantic Web framework, ontologies are best represented in the OWL. However, the limited expressiveness of OWL and its clear distinction between classes and instances make it difficult to use OWL ontologies computationally. A particular drawback is the lack of variable use on the class-level, which makes it impossible to express certain rules that would be needed to use classes to model complex structures that are prevalent in e.g. molecules or cells. Especially with structures such as molecules, there exist large numbers of real world instances that would fit the description of the molecule class, e.g. H2O. In order to express a reaction such as \(2H_2 + O_2 \leftrightarrow 2H_2O\), a SWRL rule needs to be deployed that operates on all instances of the kind. However, it is difficult to express that as well, because not all \(H_2\) or \(O_2\) in the world actually participate in the formation of water. When scientist state that “two hydrogens and one oxygen form water in a reaction”, they are generally not talking about one specific instance of that reaction. Neither are they talking about all actual instances of this reactions that currently take place. It is more the expression of a notion of this reaction.
Scientists have a mental model of their domain in which theoretical, archetypal instances of things operate that can be generalized to real occurrences.

Here, I want to clarify the notion of archetypal instances for use in life-science ontologies to allow for factual statements of general rules within the framework of OWL. An ontology that follows this paradigm is not realist in the classical sense. None of the metadata in the ontology has an actual referent in reality, but rather types of things that behave identically in identical situations. This allows not only for computational use of ontologies, but also to describe differences in behavior of the same sort of thing in a different context. E.g. an oxygen atom in $H_2O$ plays a very different role from one of the oxygen atoms in $O_2$ and should thus be represented differently. GlycO is to our knowledge the first ontology to use this paradigm of archetypal instances in a coherent way.

I argue for the legitimacy of purely conceptual instances in ontologies for information systems. There are pragmatic reasons for this, but I also believe that ontologies in information systems will, with some exceptions, necessarily be specifications of conceptualizations in Gruber’s sense (Gruber, 1993b). He elegantly sidesteps all metaphysical considerations by placing ontologies in information systems in the realm of artifacts of our conceptualization. Whereas this move by itself has ontological implications, it gives the designer and the user the greatest flexibility.
The problem of deciding where to make the cut between classes and instances and what to consider as an instance is well known in ontology design (Noy and McGuinness, 2001). Even though OntoClean (Guarino and Welty, 2002) describes some fallacies that can occur when making wrong choices for classes vs. instances, it is often seen as an arbitrary, domain- or task-dependent choice. Noy and McGuinness (2001) give a good example for the wine ontology in which the designer has to decide whether the type of wine or the single bottle are of particular interest to the users of the ontology and thus whether to design the type of wine as an instance in order to express statements such as “Marcus \(\rightarrow\).likes \(\rightarrow\).2008 Etude Wines Pinot Noir” or as a class in order to express statements such as ”Marcus \(\rightarrow\).bought \(\rightarrow\) 2008 Etude Wines Pinot Noir bottle number 123456”.

By analogy, an ontology in the glycan domain could describe individual glycan molecules. With 1015 (or more) chemically identical glycan molecules in a purified laboratory sample, this would be a tedious and useless endeavor. It makes much more sense to describe archetypal glycan molecules. Within the context of GlycO, it is not very useful to have a simple, mostly textual description of the glycan structure, as in most carbohydrate databases. To describe the complex structural features of glycans, each glycan is composed of several building block instances that model the monosaccharide residues. Each residue instance is richly described by the sub-tree it terminates and by additional properties that define how it is chemically linked to the next residue in the glycan. This level of granularity is chosen for the description because these individual features can be associated with the physiological properties of the glycan and the cellular machinery involved in its biosynthesis, catabolism, recognition, etc.

For the current version, which focuses on the N-glycans subclass, this is accomplished by defining a tree structure of archetypal residue entities that subsumes most N-glycans. That is, almost all of the known N-glycan structures can be completely specified by choosing a subset of the nodes of this tree. This subset forms a connected subtree that includes the root residue. This tree (known as GlycoTree) has been previously described in Taka-
hashi and Kato (2003), and its structure is formalized as a collection of interconnected, archetypal residue instances in GlycO. See Figure 4.3 for an image of GlycoTree.

In spite of its practicality, the use of archetypal residues to describe glycan structures evoked some ontological problems. If a glycan instance is chosen as a representative for all real glycans that have this structure, can also a residue instance that appears in many glycan instances be at that same level of abstraction or does each archetypal glycan need its “own” set of residue instances? The key question here was to which extent an instance is determined by its context. In particular, the issue was whether it was ontologically justifiable to have each residue instance determined only by its chemical structure and the residue to which it is linked in the glycan and not also by a particular type of glycan it appears in. From a purely structural point of view this was justified with the GlycoTree structure elaborated by Takahashi and Kato (2003). Practically, it is justified by the reduc-
tion in the number of residue instances that results when different glycans can “reuse” the same residue in the same position. Semantically, this decision is justified because it reflects the way glycans are actually synthesized along their metabolic pathways, where enzyme-catalyzed reactions “add” new residues as leaves to the existing glycan tree structures or “move” the entire glycan to a protein. A specific type of residue is added in a reaction catalyzed by a specific enzyme at a specific position in the precursor glycan. Therefore, even when a specific reaction is investigated, this metaphor holds true. The precursor glycan is changed by adding another residue, but the monosaccharide molecules are still the same that were in the precursor, with the addition of a new substructure. It is therefore also sensible to make semantic distinctions between residues of the same type when they are in different positions in the glycan. It is known, for example, that a mannose residue in position 1 is functionally different from a mannose residue in position 4. What remains to be demonstrated is whether residues in the same position in different glycans can be mapped to a particular function or participation in a metabolic pathway. This assumption is naturally underlying the current implementation. The chosen design can help determine whether this assumption is valid or not, because it is easily falsifiable on a case-by-case basis. For example, sets of glycans can easily be established that contain the same archetypal residue instance and can then be queried as to whether the members of the set have common biological functions or are part of the same metabolic pathway.

Another issue of granularity is deciding which granular partitions of the world are represented (Bittner and Smith, 2003). Even in the molecular context of GlycO, different levels of granularity arise, especially when it comes to the representation of chemical linkage. Conceptually, larger molecular fragments are linked together, for example in glycans that attach to proteins. However, the actual link is naturally between two atoms. Intermediate links can also be asserted, such as the link between the glycan root residue and the amino acid in the protein that it attaches to. This issue was resolved by allowing chemical links to embody all these links recursively. The link is promoted from a simple relationship
to a first class object that is defined by the two objects it links and by a more refined link.

Furthermore, atoms are parts of molecular fragments, which in turn are parts of molecules. This is an example of a partition into bona-fide versus fiat objects (Bittner and Smith, 2003). Molecules exist as wholes independently of other objects. Molecular fragments describe functional partitions, even though they actually exist as such for extremely short amounts of time during chemical reactions, and should thus rather be seen as fiat objects.

### 4.2.4 Application example

In the GlycO ontology for the complex carbohydrate domain, the notion of archetypal instances is used to model complex carbohydrates in a granular fashion. Each such instance is composed of building blocks that model simpler molecules, such as monosaccharides, and atoms that compose the simpler molecular structures. Using this modeling paradigm, complex structural and causal dependencies can be inferred between structures and processes. It also facilitates quality control. Guarino’s article on determining ontology quality (Tapsure2004ontologyevaluation) proposes comparing an ontology to a conceptualization. In most cases there are no previous conceptualizations at hand and even if there are, a mapping has to be performed. In the case of GlycO, entries in carbohydrate databases can be compared to the structures modeled in GlycO. The ontology defines a ground truth. Mapping to it is sound. If a mapping succeeds, the database entry is valid. If the mapping does not succeed, the database entry is either wrong, has some properties underspecified or describes a previously unknown structure. If the community of experts agrees on the latter, automatic mechanisms update the GlycO ontology to account for the newly recognized structure.

The interface between the ontology and the user has been developed to represent entities using the same visual metaphors that domain scientists use (Eavenson et al., 2008). This facilitates the mapping between the conceptual representation in the ontology and the actual instances in the world. Figure 4.4 shows such a depiction of a Glycosylation...
pathway.

Figure 4.2.4 shows partial snapshot of the N-Glycan synthesis pathway using the so-called Cartoonist representation (Goldberg et al., 2005) of five N-Glycans in the upper row. The lower row shows distributions of transcriptomic and glycomic data for each step of the pathway. It is clearly visible that the N-Glycan structures change by only one fragment in each step, the rest remains unchanged. Using archetypal instances as building blocks in the representations of each Glycan models this behavior. The representation of the second Glycan in this pathway has as its part exactly the same archetypal monosaccharide instances that the first one has, plus one.

### 4.2.5 Instances as Archetypes of Concepts

In OWL, this kind of metamorphosis from one molecular structure to the next in a pathway cannot be modeled on the class level. It would not be possible to make a judgement about where the monosaccharide instances in each Glycan came from. It can thus be said that even though a conceptual level is represented, modeling it in the ABox may be a better representation of reality as a model based on TBox universals could be. Practically, using archetypal instances instead of universals gives us the ability to apply path queries or apply
ABox reasoning to deductively identify implicit relationships between the archetypes.

In large bodies of uniform entities, such as molecules, it often makes sense to abstract from the individual occurrence and focus on the universal attributes. Every $H_2O$ molecule, for example, has the same chemical properties. In this case, the modeling can resort to abstract entities. Depending on the ontological commitment made, this might bear some problems. Kuśnierczyk (2006), discusses two possible readings of the term abstract entity. The first being a universal, as in (Bittner et al., 2004), the second being the concept of an entity we have in mind.

A particular bridge in Amsterdam may be a real-world entity. However, once we start discussing physical properties of that bridge, for example how much it could possible hold before it collapses, we conceptualize it. An ontology built e.g. for the purpose of capturing the statics of old roadways that has an entry for this bridge might state structural properties of it, but not who built it or how many boats pass under it day after day. The actual bridge is its entirety of being. Every representation is necessarily incomplete and depending on a convenient conceptualization. Thus, any representation of a real-world entity in an ontology will necessarily be a representation of a conceptualization. This brings us to the dilemma of explicitly or implicitly describing real-world entities in ontologies. In the semantic web context with OWL as its standard, we are limited to representations of classes and instances thereof. An abstract entity, such as my community’s concept of that bridge can then, according to the above mentioned discussion, be described as a class that has only this one bridge in its extension or as a conceptual instance.

Conversations between people often take place on a conceptual level. We talk about fictional characters; we all have different imaginations of places, that maybe not all the participants have seen or different imaginations of scientific models. It is not clear what each individual’s mental representation of these concepts are, but it seems that parts of these conceptualizations are shared, at least on a symbolic level. Sometimes there are different interchangeable models or metaphors. For example, molecules are represented as chains or
clusters of balls, as ball and stick models or as formulae according to different conventions. More complex molecules such as proteins are visualized using a ribbon representation. We imagine single atoms in the best case as a (usually solid) core with electrons either spherically moving around or located in a cloud relatively far away from the core. Harrison and Treagust (2000) studied how 11th grade high-school students mentally represent simple and more complex chemical compounds and there is no reason to believe that scientist have a fundamentally different metaphors to mentally represent these compounds. Naturally, the more expertise somebody has, the more refined the mental representation will be. However, depending on the reasoning task at hand, different mental models will be used. A quantum physicist probably deploys a more nave model when she puts a glass on a table without wondering why it does not fall through the surface. A chemist does not take into account all possible variations of isotopes in a molecule unless it is important for the task at hand. Bittner and Smith (2001) take this into account by addressing granularity.

We assume that when people reason about concepts they deploy the same kinds of mental models they would use if they were reasoning about an actual entity. We go so far as to say that reasoning never has the actual entity as a component, but always some representation of it. The outcome of the reasoning is then again mapped to the world outside the mental representation.

From this point of view it is legitimate to design an ontology that is by definition a specification of a conceptualization. It seems that from a realist point of view, mental models should be just as real as other intangibles. Given that there seems to be consensus in the scientific community about these models, the ontology becomes a specification of a shared conceptualization (Borst, 1997; Gruber, 1993a). The danger is in crossing the boundaries. An ontology with instances representing mental models of a thing should not contain instances that refer to the real world representations of these archetypes.

In this modeling paradigm, it is important to be careful not to mix references to real-world instances and archetypal concept instances. The scientific model is formally repre-
presented, not the world it is trying to describe. The great advantage is that no claims have to be made about its correspondence to actual states of the world. The disadvantage is that we no claims about this correspondence can be made. However, it seems epistemologically more honest to leave the interpretation up to humans who can draw the connection with the real world.

### 4.2.6 Implications

Deeply entrenched with the realist philosophy is the correspondence theory of truth. This normative theory describes an idealized situation that requires a full and unhindered access to the world. From the point of view of information systems, reasoning happens on the basis of a formalization of the world, a digital metaphor built for symbol processors. The outcome of automated reasoning is always limited by the accuracy of the model. An ontology as a specification of a shared conceptualization thus assumes a coherence and/or consensus theory of truth. Even if there is a detectable correspondence between the world and the ontology, the information system will judge its reasoning by coherence. A good synthesis is given by Lakoff and Johnson (1980) in their experientialist account of truth, which incorporates elements of correspondence theory, coherence theory, pragmatic theory and classical realism: We understand a statement as being true in a given situation when our understanding of the statement fits our understanding of the situation closely enough for our purposes.

### 4.3 Populating the Ontology

#### 4.3.1 General Considerations

Creating ontologies is usually costly. In addition to a schema design, the actual domain knowledge in form of instances needs to be gathered, conceptualized and formalized. CYC
(Lenat and Guha, 1989) and GO are examples of ontologies that require high maintenance, due to the need for manual curation. This is not an issue in ontologies that only describe a schema to be used for database integration or as vocabularies. But since instance descriptions in GlycO are very different from those found in databases, ways to automate this process needed to be found. The objective in the development of GlycO was to have an expressive and restrictive schema that allows automatic and hence less expensive maintenance, given that semi-structured and reliable information is available for its population.

4.3.2 Populating GlycO from trusted sources

With CarbBank (Doubet and Albersheim, 1992), KEGG (Ogata et al., 1999) and SweetDB (Loß et al., 2002), several databases exist that contain trusted and up-to-date information about glycan structures. Even though CarbBank was discontinued, its content is of high quality and it is still used as a reference in other databases. The GlycO schema specifies more complex relationships than these databases. A large number of properties not specified in their schema can be computationally inferred from the information given in the databases and are then explicitly added to the glycan description in the ontology. Hence these sources are used to populate the ontology with carbohydrate instances, alongside other sources for the population of gene and protein information. Whereas each of the databases can contain incorrect entries, it is less likely that all three have the same incorrect entry. For this reason information is extracted from all these databases and compared with the canonical GlycoTree representation in the ontology during the population. To gather the data, the Semagix Freedom toolkit (Sheth et al., 2002) was used that facilitates extraction of information from semi-structured websites and converts it to a structured representation that can be exported as XML or RDF or accessed via an API.
4.3.3 An Intelligent Population Algorithm

A structured representation of data does not necessarily guarantee its usefulness. Since the information was extracted from different sources, it has to be disambiguated to avoid having differently named copies of the same structure. As mentioned above, a simple textual description of structures is not suitable for our purposes and would only give an RDF encoding of already existing databases. In order to disambiguate the potential instances, the textual description of the structure was converted into the internal GlycoTree representation. This was performed using a multi-step process in which ambiguity is progressively removed as more meaningful representations are generated.

Conventionally, glycans are represented in the so-called IUPAC format, which is a two-dimensional textual representation that visually reflects the inherent tree structure and is easily comprehended by the human eye. Unfortunately, this representation is not unique. A web service is provided that converts this representation into the structurally unambiguous Linear Notation for Unique description of Carbohydrate Sequences (LINUCS) (Bohne-Lang et al., 2001). Since this conversion is purely based on structure, it does not disambiguate different naming conventions for the substructures of the complex carbohydrate, the monosaccharide residues. For this purpose, another conversion is used that transforms the LINUCS representation into the XML-based GLYcan Data Exchange (GLYDE) format (Sahoo et al., 2005), which semantically disambiguates the different naming conventions of monosaccharide residues. XML has an inherent tree structure and GLYDE uses this fact. A child monosaccharide residue in a glycan is simply represented as a child node in the XML representation. This makes it relatively easy to perform tree operations on this representation. (See Figure 4.5 for the population workflow)

In the GlycoTree model each monosaccharide residue is defined by its type, its linkage and its position in the GlycoTree. Because of its archetypal representation, the root node of a glycan can potentially be the root node of any sub tree of the GlycoTree. The population
algorithm identifies and assigns the sub tree that corresponds to a particular glycan that is to be instantiated in the ontology. This is done by looking for sub tree isomorphisms. Several efficient sub tree isomorphism algorithms are available (Raymond and Willett, 2002). In our case, because of comparable small glycan structures, a depth-first search was sufficient. Additionally, the glycan constitutes a complete sub tree isomorphism; i.e. there cannot be a node in the glycan representation that is not part of the larger tree, nor can there be merely a homomorphism such that edges in the GlycoTree would need to be contracted to accommodate the glycan structure. If no isomorphism can be found, new GlycoTree nodes are generated automatically to complete the ontology. Here as well a report is generated so the domain expert can verify the correctness. New tree nodes can be inappropriately generated as a result of an incorrect structural description or classification of the glycan in the database. Several incorrect glycan descriptions were identified by checking all new nodes that were generated during the population process. As only a few new nodes were generated, this is much easier than checking the entire set of glycan instances for errors.

The population algorithm will also be used to automatically build minimal trees for other glycan subclasses, such as O-glycans and glycolipids, which have not been classified entirely in such a tree structure. In (Hashimoto et al., 2004) such tree structures are built, but
only cover 61.2% of the known carbohydrate structures. The set of GlycoTree nodes that represent a particular glycan can be easily compared to another set of nodes that represents a different glycan instance in the ontology. Two glycans are the same if and only if their tree node sets are identical. This method of disambiguation proved to be the more robust than other criteria, such as a common identifier, which is unreliable because every database uses proprietary accession numbers. Although all of the databases that were used as trusted sources make reference to CarbBank identifiers, CarbBank is no longer actively curated and these databases contain glycans that do not have a CarbBank ID.

4.4 Evaluation

It is difficult to measure the quality of an ontology. Guarino (Guarino, 2004) proposed an evaluation based on precision and recall with respect to a reference conceptualization. This of course requires a formal conceptualization that applies to the same domain or of a meta domain. With respect to the OntoClean ontology, for example, such a formal evaluation can show whether certain meta-properties of concepts are correctly assigned in the ontology. GlycO follows this meta-methodology. It is less common to find an applicable conceptualization of the same domain, because new ontologies are usually created because there is no ontology for the domain readily available.

Another dimension for evaluation are structural metrics that assign numerical values to criteria such as depth, breadth, fan-outness, etc. (Gangemi et al., 2005; Tartir et al., 2005). These metrics are useful especially in large ontologies to get an idea of their structural character. Of course, none of these metrics can really tell us how useful an ontology will be and how well it models its domain. Table 4.1 shows the results of comparing GlycO to other biomedical ontologies using these metrics. Instance information is not taken into consideration. GlycO shows the highest connectivity, indicating a rich set of well defined and logically restricted relationships. The average number of sub terms gives an indication
of the fan-out, but also the depth of GlycO. In a comparable fan-out measure, when siblings are counted, the number of siblings ranges between 1 and 15 with an average of 6.

Table 4.1: Comparison of GlycO to other biomedical ontologies

<table>
<thead>
<tr>
<th>Ontology</th>
<th>No. of Terms</th>
<th>Avg. sub-terms</th>
<th>Connectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlycO</td>
<td>324</td>
<td>2.5</td>
<td>1.7</td>
</tr>
<tr>
<td>ProPreO</td>
<td>244</td>
<td>3.2</td>
<td>1.1</td>
</tr>
<tr>
<td>MGED</td>
<td>228</td>
<td>5.1</td>
<td>0.33</td>
</tr>
<tr>
<td>Biological Imaging methods</td>
<td>260</td>
<td>5.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Protein-protein interaction</td>
<td>195</td>
<td>4.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Physico-chemical process</td>
<td>550</td>
<td>2.7</td>
<td>1.3</td>
</tr>
<tr>
<td>BRENDA</td>
<td>2,222</td>
<td>3.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Human disease</td>
<td>19,137</td>
<td>5.5</td>
<td>1.0</td>
</tr>
<tr>
<td>GO</td>
<td>200,002</td>
<td>4.1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Metabolic pathways can be queried using GlycO, even though they are not explicitly defined the way they are in some databases. A metabolic pathway is essentially a sequence of reactions that lead from one chemical compound to another. The advantage of our representation is that any path between compounds can be shown by traversing relationships, even if these compounds are not explicitly assigned to a specific pathway, given that all the reactions that are involved are formalized in the ontology. This makes the representation of pathways in the ontology more flexible than that in many databases. Figure 4.6 shows the GlycO representation of some steps in the N-Glycan biosynthesis pathway.

Another application that requires sophisticated algorithms on databases is described in (Hashimoto et al., 2004). The different glycan trees that the authors identify are inherently encoded in the archetypal residues and links and can thus easily be queried as well as visualized.
Figure 4.6: A part of the N-Glycan biosynthesis pathway as encoded in GlycO. For better visibility, only few relationship types are visualized. \textit{N-glycan\textsubscript{b-D-Glc}pNAc\textsubscript{13}} is the beta-D-Glc\textsubscript{p}NAc residue number 13 as enumerated in the GlycoTree model.

4.5 Conclusion

GlycO is not only a vocabulary or a schema meant for database integration, but provides a rich description of the knowledge in the glycoproteomics domain, semantically describing interactions and functions of structures and their substructures as well as their synthesis.

In the context of this modeling, mereological problems were encountered and addressed. By promoting some of the relationships in the ontology to first class objects, recursive definitions of these relationships allow their expression on different levels of granularity.

By semantically modeling the structure of molecules with reusable archetypal instances, the hypothesis that larger structures exhibit properties and functions that can partially be inferred from the knowledge of the properties and functions of their substructures can be evaluated. The GlycO schema allows a glycan structure to be represented as more than the sum of its parts, paving the way for the identification of the molecular basis for
emergent properties. To our knowledge is GlycO the first ontology that models its domain in such detail as described. The formalization of this knowledge allows immediate access to information that so far is only available through specialized tools and algorithms that work on the textual representation in the various biochemistry databases. It was shown that with a sufficiently rich schema alongside trusted sources, automatic extraction, modeling and classification of high-quality instance data is possible.

The creation of the GlycO ontology implemented the *Circle of Knowledge* insofar as the ontology draws from explicit background knowledge that was available in existing ontologies to a small extent, but mostly from the tacit knowledge that is held by domain experts. The T-Box, the Archetypal instances and the glycan instances that were acquired using the automated extraction procedure are automatically added to the ontology, which in turn becomes part of the background knowledge. Glycan instances are added automatically, when their structure is entailed by the GlycoTree model, which functions as a justification-by-coherence means of justification and truth-finding. In case an extracted structure is not entailed, it is presented to domain experts to confirm that it is incorrect. In the rare case that a structure could be found that is correct even though it is not entailed by the GlycoTree model, GlycoTree can automatically be augmented.
Automatic Domain Model Extraction

Models are to be used, not believed.

(Henri Theil, Principles of Econometrics)

This chapter describes the Information Extraction aspect of the Knowledge Acquisition cycle of this work. Since IR applications are the driving force behind the validation of extracted statements, the use cases and examples are mostly build around such tasks.

Conceptual search, browsing and classification of documents using background knowledge from domain models has been the topic of extensive research in the Semantic Web and IR communities (Castells et al., 2007; Vallet et al., 2005; Mayfield and Finin, 2003). A wider adoption of these methods is largely dependent on the availability of personalized and focused semantic domain models or ontologies. However, ontologies, domain models, controlled domain vocabularies and taxonomies that can be used for annotation and retrieval tasks are either not available at all or expensive to obtain. The Taxonomy warehouse portal\(^1\), for example, offers a variety of taxonomies for purchase. Apart from the associated cost that can often exceed tens of thousands of dollars per taxonomy, these vocabularies are still static and topic-centered instead of being

\(^1\)http://www.taxonomywarehouse.com
ing dynamically developed for a specific user’s interests at a particular point in time. An automated solution that can deliver personalized models on-demand in an affordable and scalable manner is thus highly desirable.

The task of ontology learning (Maedche and Staab, 2001) is concerned with automatically extracting concepts and relationships from an information corpus. This has often been done by parsing sentences and identifying relevant noun phrases that act as concept descriptors and verb phrases that identify pertinent relationships, thus extracting a graph of concept and relationship mentions. Ontologies, however, are by definition more than collections of identifiers and syntactic predicates that are used within a corpus. An ontology, often defined as “formal, explicit specification of a shared conceptualization” (Gruber, 1993a), needs to formally define a domain in terms of concepts and relationships. It should also reflect an agreement among those who hold knowledge in the domain and who want to use the ontology as a means to communicate thoughts and facts about the domain.

Automated methods for ontology learning generally lack the capability of properly identifying concepts beyond extracting named entities. This is due to the fact that algorithms lack the ability to properly abstract from concept mentions/identifiers to concepts and to abstract from syntactic to semantic predicates. This means a bottom-up approach to concept and relationship identification will often end up with merely noun phrases and verbs.

One hindrance for ontology learning applications has often been that complete correctness can not be guaranteed in automated methods. Some of the strict requirements for ontology learning can be relaxed when creating domain models for IR applications (Gulla et al., 2007). It is often not necessary to have ontologies that are formally consistent and only contain true knowledge. In IR tasks, users are generally content with results that are mostly correct, but not necessarily flawless, because a list of web results is quickly scanned and irrelevant results are simply not looked at. Users are more interested in finding the best possible results amongst the first ones that are displayed. A filter that improves search
and/or classification results is thus beneficial, even if it contains some concepts that arguably do not belong to the domain of interest or contains some incorrectly extracted facts and thus may cause some filtering mistakes. Because of the potential errors in automatically created models, I will refrain from using the term ontology and use the term domain model instead.

Even if for models used in IR the requirement of absolute factual correctness can be relaxed, it is still desirable to have a consistent and grounded representation of concepts and relationships. In faceted browsing, for example, facets should be treated or displayed according to the semantics of the concepts or relationships that create the facet. Models that are created solely based on elevating syntactic structures to semantic predicates cannot achieve this. Facets may be ambiguous or several facets created by different predicates may have the same semantics.

Community-created or peer reviewed fact-corpora such as DBPedia (Bizer et al., 2009b) or UMLS provide a vast amount of well-defined concepts and relationship types, which makes them ideal candidates for a top-down extraction of conceptual knowledge. However, whereas these corpora have a large coverage of their domains in terms of concepts, they are very sparsely populated with facts that instantiate the relationship types. For example, DBPedia 3.6 has about 7 million asserted facts involving named object properties other than category membership and type, i.e. subject and object of the triple refer to a URI-resource. Given that Wikipedia currently has about 3.7 million articles that describe concepts, this means that every resource has on average less than two connections to other resources. Since Wikipedia and DBPedia are growing corpora, the fact coverage may become denser over time, but for a domain model extraction application, it will likely be too sparse for some time to come. Hence it is necessary to have a bottom-up approach to automated fact extraction from free text that operates on the concepts and relationship types that were identified top-down.

When information sources are chosen for the information extraction task that take
the social character of knowledge aggregation into consideration, agreement is practically built into the ontology learning task. This assumption allows us to extract domain models that, despite their focus on an individual’s interest, represent a shared understanding of the knowledge that underlies the model creation, thus fulfilling one of Gruber’s requirements for formal ontologies (Gruber, 1993a). Recent research in the realm of Web 2.0 also emphasized the communal creation of information and the common designation of concepts, for example in Wikipedia (Hepp et al., 2007) or social bookmarking sites (Halpin et al., 2007) and projects that take advantage of community-assigned designators to create more formal representations, such as DBPedia (Bizer et al., 2009b).

Based on this line of reasoning, the following requirements are identified for a domain model extraction framework that can feasibly be used for IR applications:

1. Identification of concepts that are pertinent to a domain of interest

2. Extraction of relationships between the identified concepts

3. Efficient extraction to reduce delays in IR applications

In this chapter I present Doozer++, an approach to on-demand creation domain models that approximate the information seeker’s intent and context of inquiry using top-down concept identification and bottom-up fact extraction. The domain model extraction framework builds models representing a domain or a context based on keyword descriptions or queries. It builds a concept hierarchy that delineates the domain of interest and embelishes it with automatically extracted facts pertaining to its concepts to facilitate contextual browsing or faceted exploration of the search space.

The technical contributions of this work are:

1. A domain hierarchy extractor that creates a Domain Definition from a conceptual corpus, such as Wikipedia.

2. A concept-centric, distantly supervised relational-targeting semi-open IE algorithm
using surface patterns occurring in free text for the extraction of the Domain Description.

3. A statistical pertinence measure that facilitates dealing with semantically overlapping types of relationships in a dynamic, unsupervised fashion.


5. Recall enhancement using a pattern generalization algorithm.

6. Extensive qualitative and quantitative evaluation of the Domain Description part, including analysis of the extracted patterns.

The framework is evaluated extensively, first by separately evaluating domain definition and domain description before evaluating the combination of both.

To give an idea of the system’s utility in creating models for guided browsing, the development of a model created for the area of human cognitive performance is demonstrated. The model provides browsing background knowledge to Scooner (Cameron et al., 2010), a semantic browser that allows browsing along semantic trails. An excerpt of the model is shown in figure 5.1. The full model contains hundreds of classes and entities from the cognitive science domain.

The remainder of this chapter is structured as follows. Section 5.1 gives a broad overview over the steps involved in the domain model creation. Section 5.2 places this work in the context of related work and familiar approaches. The Domain Definition step is discussed in Section 5.3, followed by the fact extraction for Domain Description in Section 5.4. Section 5.5 discusses the combination of hierarchy creation and information extraction. In Section 5.6, the approach is evaluated extensively using manual and automated qualitative and quantitative methods.
5.1 Background

This section summarizes conceptual considerations that went into the 2-step approach of (a) Domain Definition and (b) Domain Description to model creation. It will discuss semantic and IE-related issues involved in model creation as well as the necessary assumptions made about the data in order to overcome these issues.

5.1.1 Domain Definition

Domain Definition is accomplished by restricting existing structured or semi-structured sources to only contain concepts pertinent to a focus domain. In our work we use Wikipedia as a knowledge source. We make the assumption that most concepts and entities of interest are represented by articles in Wikipedia and concept labels are represented by article titles, titles of redirect pages and anchor texts that link to the articles.

Over the years, Wikipedia has become a high-quality encyclopedia. With an ever growing number of articles, Wikipedia covers an impressive number of concepts of general interest. A Nature article from 2005 found that the amount of factual errors in Wikipedia
is not significantly different from those in the Encyclopedia Britannica (Giles, 2005). This surprisingly high quality of a community-created encyclopedia may be explained by “The Wisdom of the Crowds” (Surowiecki, 2005). According to this theory, large numbers of people are able to solve difficult problems, as long as they are independent, given a good infrastructure and their answers are aggregated in an intelligent manner (Thomas and Sheth, 2011). Wikipedia allows authors full independence and the ability to change (almost) every article, regardless of the author’s credentials. The underlying assumption is that the community will correct the mistakes of its single members. In previous work (Thomas and Sheth, 2007) we demonstrated how most articles on Wikipedia evolve over time and converge to a stable state. This work can also be used to filter out articles that are not yet mature and can thus ensure that information is only extracted from high-quality articles.

Even more than the actual content of the articles, the naming and classification of concepts is of particular interest to the hierarchy creation part of this work. It has been shown (Hepp et al., 2007), that the URLs of Wikipedia articles are valuable as general concept identifiers in ontologies. The main advantage of using these identifiers is that they have been community-vetted and are unambiguous.

Taking these findings into account we developed the on-demand domain hierarchy creation application Doozer (Thomas et al., 2008), which is also detailed in section 5.3. Thereby, a hierarchy of concepts that are pertinent to a user-envisioned domain of interest is automatically carved out of the Wikipedia article- and category graph. This extracted hierarchy describes a domain of interest as specified by a user in a keyword description. This description can be fairly elaborate in the form of a long boolean query, but it can also resemble a simple query that one would send to a Web search engine. In an interactive version of the model creation application, the user can refine the model creation by rewriting the query or otherwise extending or reducing the scope of the model.

The task here is to extract a domain definition that clearly focuses on user-interests. This process follows an “expand and reduce” paradigm that allows us to first explore and
exploit the concept space before reducing the concepts that were initially deemed interesting to those that are closest to the actual domain of interest.

A valid criticism of using Wikipedia as a source of conceptual knowledge is that Wikipedia does not provide enough depth of coverage for highly specialized domain ontologies. This is then reflected in a lack of concept identifiers in the extracted hierarchies. However, we believe that the benefit we get from having accurate concepts outweighs the lack of recall in some generated models. In order to expand the set of concepts available to an automatic hierarchy creation beyond the identifiers available in a curated corpus, error-prone techniques, such as Named Entity Recognition (NER) have to be deployed. NER techniques do not only face the problem of recognizing incorrect entities, most methods only recognize entities of a limited number of classes, such as Person, Organization, Gene, Disease or Event. In future work we want to expand Doozier in this direction, but acknowledge the difficulties that come with such a step. However, prior knowledge about types improves NER significantly (Ratinov and Roth, 2009) and the amount of background knowledge available on curated corpora such as Wikipedia and UMLS can be used for this task.

Given the modular nature of our work, it is also possible to create a domain hierarchy with other tools, create it manually or expand a hierarchy that was extracted with additional concepts before proceeding to the Domain Description step.

5.1.2 Domain Description

The second step, Domain Description, then embellishes the domain concepts with attributes and relationships using Information Extraction (IE) techniques.

Many approaches to IE extract assertions from text by first parsing sentences and next promoting syntactic Subject-Predicate-Object structures to semantic assertions (in Semantic Web contexts usually referred to as triples). Open IE approaches of this kind have been termed Structural Targeting Open IE (Banko et al., 2008; Wu and Weld, 2010). This
promotion of the individual occurrence of two entity mentions with a verb to a semantic predicate brings about many problems. When an assertion is extracted from a single phrase found in a single document, there is no notion of a shared understanding. Nevertheless, some extractors treat single evidence as sufficient. Some extractors, such as Textrunner (Banko et al., 2008), combat this problem by ranking extracted assertions by the number of occurrences found. A further problem is that in natural language use, there is an many-to-many relation between verbs and formal relationships. For example, the predicate $X$ causes $Y$ could be expressed in the phrases “$X$ causes $Y$”, “$X$ induces $Y$”, “$Y$ is caused by $X$”, and many more. On the other hand, the highly ambiguous verb “break” for example, can be used, amongst others, in the senses separate, interrupt, violate or destroy. Without further disambiguation, Structural Targeting Open IE is likely to interpret the verb “break” as having the same meaning in the sentences “Marcus broke the glass” and “Marcus broke the law”. These ambiguities are usually not considered in the evaluation of Structural Targeting Open IE systems. Instead, it is often evaluated how well a sentence was dissected into its syntactic subject object and predicate.

This ambiguity of natural language leads to many verbs expressing the same property on the one hand and ambiguous verbs that express several properties on the other. Finally, some relationships are rarely expressed as verbs. For example, we do not say or write “Ottawa capitalizes Canada”. Rather, we will find phrases such as “A magnitude 5.5 earthquake recorded near the Canadian capital of Ottawa has rattled some nerves in the Chicago area.”

**Relational-Targeting Open IE**

We address this ambiguity problem by using supervised pattern-based extraction rather than NLP-based extraction. Relationship types are manifested in many different patterns across a corpus. The domain description algorithm described in this chapter learns these patterns by finding examples for triples from a fact corpus in text and then generalizing the
patterns found for single facts to patterns for types of relationships. Fact corpora are widely available these days on the Linked Open Data (LoD) cloud. This kind of Information Extraction has been termed Relational-targeting open IE (Wu and Weld, 2010). In our case, rather than assuming a completely open number of relationships, the openness of the extraction stems from the openness of the fact corpora on LoD, which are continuously evolving and the algorithm continuously trains on the types of relationships that emerge. Hence we refer to our approach as semi-open Relational-targeting IE.

Semi-open Relational-targeting IE resolves relational ambiguities in the training phase by accumulating different patterns for a semantic relationship. The problem that a pattern such as “X broke Y” can indicate different semantic relations still persists. However, thanks to the training procedure, we have an idea how likely it is that it indicates e.g. violate or destroy. In the application phase, evidence for a relationship is not just taken from one occurrence of a pattern, but the accumulated evidence of multiple patterns between the surface representations of two concepts. Whereas a single occurrence of “Marcus broke the glass” is ambiguous for the extractor, the accumulated evidence of “Marcus broke the glass”, “Marcus shattered the glass” and “Marcus destroyed the glass” disambiguates the relationship between Marcus and the glass.

In many cases, structural targeting open IE also does not enforce the presence of meaningful concept or entity descriptors in the subject or object of the extracted statements. Many of the extracted statements by ReVerb (Fader et al., 2011) or WOE (Wu and Weld, 2010), for example, are pronouns (e.g. you, she, we, her, it, etc.), or phrases that are only referring to concepts in context (e.g. “six of 11 countries” or “the legislation”). Not only does this result in meaningless extracted statements, it can also lead to adding incorrect statements to proper concepts. For example, the noun phrase “the legislation” was extracted as a subject in the triple “the legislation → achieved → a number of needed reforms”. In

\[\text{A collection of triples extracted by TextRunner, WOE and Reverb and their evaluations can be found here: }
\text{http://www.cs.washington.edu/homes/ afader/data/reverb_emnlp2011_data.tar.gz}\]
the original sentence, “the legislation” is a synecdoche (i.e. a deferred reference) for a particular legislation that was probably introduced in a previous sentence. However, taken in isolation, the triple may be seen as referring to the concept of legislation in general. This makes obvious the need for enforcing concept reference in a concept-centric extraction to ensure the validity and proper reference of a triple’s subject and object.

Relational targeting open IE is, in our opinion, superior to structural targeting for the task of creating formal domain models, because it allows mapping of extracted relationships to existing schemas. Instead of elevating syntactic predicates to semantic relationships, formally defined properties are represented by a collection of patterns that have been identified as representing the relationship in a distantly supervised training procedure. It uses efficient surface patterns to identify and extract relationships (Agichtein and Gravano, 2000; Turney, 2006). Not only can expensive POS tagging and parsing steps be skipped, but index representations of textual data can directly be used. Search engines already represent web site data in the form of term vectors and position vectors. Thus, patterns and entities can be identified in the index itself, rather than having to analyze the full text.

The Many-to-Many Challenge

A further challenge that a supervised open IE method has to deal with is analogous to the one described for the NLP-based open IE methods. Many types of relationships are semantically overlapping or one type of relationship is entailed by another. The overlap can be intensional or extensional.

Intensional overlap is given when a property is a subproperty of another and all instances of the subproperty are necessarily instances of the superproperty. For example, the relationship physical part of entails part of. Intensional equivalence is given when the definition of both properties is exactly the same.

Extensional overlap is given when the instances of two relationships overlap without
necessarily belonging to both. For example, the relationships *birthplace* and *deathplace* often share common subject-object pairs, because e.g. a person was born, lived and died in the same place. Extensional similarity is given when the instances of relationships happen to overlap, but the definitions of the relationships are different.

In order to improve extraction in light of intensional or extensional overlap, the algorithm should be able to find distinctive features of extensionally overlapping relationships, even though many training examples are overlapping and thus the pattern manifestation of the relationships is similar. It also needs to find features that intensionally overlapping relationships share, so they are not seen as discriminating features between these relationships. These classification problems inherent to OpenIE are addressed by the development of a semantic pertinence measure for intensional overlap and an entropy measure for extensional overlap. The entropy measure identifies patterns that are highly indicative of particular types of relationships whereas the pertinence measure guarantees that same patterns for intensionally overlapping relationships are not penalized.

Each relationships will be represented by the collection of patterns that were found as the manifestation of the relationship in text. In particular, each pattern is assigned a probability of being an indicator of a relationship. Thus, a vector space representation of these probabilities is an intuitive choice. In such a representation, *a relationship is expressed as a vector of pattern probabilities* and the collection of relationships is expressed in a *relationship-to-pattern (R2P) probability matrix*. Vector space representations have been proven to be successful in document classification and relationships extraction (Turney and Pantel, 2010).

Based on these considerations Domain Description is accomplished within the following framework:

- Pattern-based representation of relationship types
- Vector-space model for the representation of relationships and patterns
• Probabilistic multi-class classifier for fact extraction
  – white box approach for better control and better analysis
  – positive-only classification
• Pertinence-based and entropy-based computation of pattern importance

5.2 Related Work

This section introduces relevant projects that either influenced the research described in this chapter or offer alternative approaches to information extraction and ontology learning. Since this project uses a two-fold approach, this section is split into subsections that introduce prior work related to the overall goal of domain model extraction as well as the individual parts of hierarchy extraction and named relationship extraction.

5.2.1 Ontology Learning

A large body of work is dedicated to the automatic creation of taxonomies or ontologies from text. Maedche and Staab have an extensive survey of methods in (Maedche and Staab, 2001). Most research work is concerned with the extraction of taxonomies from free text, where the scope of the domain is given by the scope of the text corpus, mostly focusing on combining linguistic analysis with statistical methods and formal concept analysis, see (Cimiano et al., 2004, 2005). The same group also recognized the use of automatically generated ontologies for clustering (Bloehdorn et al., 2006). A well-known application in automatic Ontology creation is Text2Onto (Cimiano and Völker, 2005), which creates models from free text guided by interaction with the user. It uses fixed patterns for the extraction of hierarchical relations and shallow parsing based open IE techniques to extract named relationships.
Most of these ontology generation efforts extract both domain definition and domain description bottom-up from text. Doozer++ bypasses the problems that arise because of syntactic and semantic ambiguities in identifying by taking advantage of top-down analysis of community generated or peer reviewed corpus that is free of ambiguities in its graph structure. The use of Wikipedia article names as universal concept identifiers has been discussed in (Hepp et al., 2007).

5.2.2 Top-Down Extraction of Knowledge

Research efforts that have made use of the Wikipedia corpus to infer taxonomic knowledge include (Ponzetto and Strube, 2007) and (Zirn et al., 2008). These efforts use Hearst-style patterns and heuristics based on Wikipedia naming conventions to identify those inter-category relationships in the Wikipedia hierarchy that are actually is-a relationships and are helpful in distinguishing between classes and instances. Another example is YAGO (Suchanek et al., 2008) that improves the classification of Wikipedia concepts by mapping them to WordNet (Fellbaum, 1998). For an application that aims at producing formally correct ontologies, these and comparable efforts constitute important preprocessing steps. However, none of those works is concerned with the restriction of the Wikipedia corpus to a specific domain of interest.

5.2.3 Bottom-up Extraction of Knowledge

5.2.3.1 Domain-Taxonomy extraction

In the Taxaminer project (Kashyap et al., 2005) taxonomies were extracted from biomedical documents with no structural knowledge of the domain available to the system. The resulting hierarchy was generated solely by identifying cohesive clusters in a hierarchy that was an artifact of a bisecting k-Means clustering process. The clusters with the highest
information gain became representative of concepts and the most salient terms in the clusters were used as concept labels. Even though the clusters were of high quality based on IR measures, the resulting hierarchy did not reflect the shared human conceptualization of the domain when it was compared to MeSH. The intuitive explanation for this is that research papers are problem-centric, i.e. the view that is taken of the domain emanates from the problem at hand. A taxonomy is world-view dependent, i.e. it tries to summarize the entirety of concepts that are pertinent to a field of interest.

A large scale taxonomy induction using Hearst-style patterns was done by Sánchez and Moreno (2008). Domain taxonomies were extracted by finding instances of these patterns in large numbers of web pages. The created taxonomies are of high quality, which can be attributed to the high-precision nature of Hearst-patterns. However, the amount of pages that has to be crawled and the resulting extraction time is prohibitive for on-the-fly creation of taxonomies.

A fully automated approach to extracting domain taxonomies from free text was proposed by Navigli et al. (2011). Given a set of seed concepts that describe the upper level of the taxonomy, the algorithm will extract a domain taxonomy from domain dependent text corpus using term extractors, Hearst-style patterns and a graph-based pruning technique. The precision for a scientific domain was 81.5%. We naturally achieve a higher precision, because taxonomic relationships are already available in the corpus for the Domain Definition. However, this work is a promising contribution that could replace or augment the current Domain Definition step in order to create deeper and more finely grained domain taxonomies. The highest extraction quality is will likely be achieved when the recognition or extraction of entities is also paired with the extraction of facts or relationships, as in Yu et al. (2011).

The clustering-based approach to taxonomy extraction shows that the way concepts are represented in text does not generally give us an indication of their taxonomic categorization. The pattern-based approaches suffer from different shortcomings. Hearst-style
patterns often indicate is a relationships, but not always. Consider, for example, the pattern “(NP) and other (NP)”. A web search for “. and other important issues/topics/matters”, for example, would suggest that e.g. *Lady Gaga is a important issue*.

Acknowledging the importance of a proper classification of concepts and the difficulties involved in bottom-up extraction of taxonomies spawned the decision to extract taxonomies top-down, from existing conceptual sources.

### 5.2.3.2 Relationship Extraction

Most bottom-up relationship extraction work uses Information Extraction techniques. Figure 5.2 puts our work in context with the current state of the art in Information Extraction wrt. the dimensions *NLP complexity* and *openness of extraction*. NLP complexity refers to the amount of text processing that was performed, e.g. POS tagging, parsing, etc. Openness of extraction refers the number of relationship types that can be extracted, from very few dedicated types, such as hyponym/hyponym relationships to an unbounded number of relationship types. *Doozer++* is placed in the no-NLP/many relationships corner. Most works have either been restricted in the number of relationship classes that are extracted (Agichtein and Gravano, 2000; Chklovski and Pantel, 2004; Hearst, 1992; Pasca et al., 2006; Snow et al., 2004) and/or have made use of parsing or POS-tagging to improve precision and recall of the methods (Banko et al., 2008; Carlson et al., 2010b; Mintz et al., 2009; Ramakrishnan et al., 2006; Suchanek et al., 2006; Weld et al., 2009; Wu and Weld, 2010).

Pattern-based information extraction has been successfully applied in past research. Hearst (Hearst, 1992) used manually identified patterns that indicate hyponym relationships. Snow et al. (Snow et al., 2004) use automatically identified pattern vectors for the same task.

A pioneer in the area of surface pattern-based extraction for general named relations
from raw text is the Snowball system (Agichtein and Gravano, 2000). It uses a pattern-definition similar to ours. The main difference is that Snowball works on a very restricted set of relations and assumes a previous named entity recognition to restrict the extraction to \(<\text{Organization},\text{Location}>\) pairs. Pasca et al. (Pasca et al., 2006) also heavily restrict the types of relationships by extracting only date-of-birth attributes from Web documents, thus achieving a very high average precision of 93.17%. Whether the system has the potential to scale up to an arbitrary number of property types is not discussed. The pattern-based approach taken in the Doozer++ system is inspired by Turney’s (Turney, 2006) work on identifying analogous word pairs. Similar to our work, Turney uses vector space representations of surface patterns without parsing or POS-tagging the text.

*Structural Targeting open IE* approaches that are not restricted to a fixed number of relations are TextRunner (Banko et al., 2008) and Ramakrishnan’s work on extracting relationships in the BioMedical domain (Ramakrishnan et al., 2006). These systems assume
that the syntactic predicate in the sentence expresses the relationship of interest, which is not always the case. Systems that use a relational targeting IE approach in conjunction with NLP techniques are Kylin (Weld et al., 2009) and WOE (Wu and Weld, 2010) which use Wikipedia Infoboxes to train Conditional Random Fields (CRFs). WOE outperforms TextRunner when using dependency parses for extracting new relationships. Mintz et al. (Mintz et al., 2009) used both lexical features as well as syntactic features. A comparison showed that a combination of both performs best, especially in higher-recall scenarios. However, even the lexical features go beyond mere surface patterns and contain POS information. Moreover, a named entity tagger identifies a limited number of entity tags, such as Person, Organization and Location. In this dissertation it is assumed that neither POS tags nor syntactic features will be available at runtime. LEILA (Suchanek et al., 2006) performs a strong linguistic analysis of the corpus and identifies complex patterns along paths through the parse tree. LEILA can also take advantage of Wikipedia-specific features, such as common page structuring, category-assignment, titles, headings, links and InfoBoxes. Here, relationship extraction is kept independent of a particularly structured corpus for two reasons. First, it should be possible to extract domain-specific relationships from any specialized corpus and second, the base for the extraction should be expanded across corpora, i.e. potentially search for evidence on the whole Web.

NELL (Carlson et al., 2010a) is a so-called never ending learning system that extracts new relationships by continuously reading the Web. It uses multiple complementary approaches to achieve better precision by coupling different learning functions. NELL is part of the Read-the-Web project (Carlson et al., 2010b), which also uses background knowledge in the form of ontologies or domain models to improve the classification of entities and relationships. Different from our work, it tries to learn new concepts using NER techniques, which leads to many incorrectly identified concepts. NELL, as well as some of the other above mentioned approaches, could be used as the domain description part of our system. However, all of these systems have some shortcomings in the context of domain model
creation, i.e. some are not concept-aware and/or not aware of proper types of relationships, which are beneficial, if not essential, requirements for domain model creation.

Other research projects have used both statistical methods and reasoning using background knowledge. SOFIE (Suchanek et al., 2009) performs a MAX-SAT test on extracted statements to see how well they fit into the KB and whether other statements can be derived. (IJntema et al., 2012) developed Lexico-Semantic patterns to achieve better extraction by incorporating semantic types right into the patterns. Defining rules and Lexico-Semantic patterns is done manually and is hence a cumbersome process. These methods work best for maintaining KBs for predefined domains in which extraction precision is very important, rather than the open domains envisioned for Doozer++.

5.2.3.3 Positive-only Classification

Similar to Wang et al. in PORE (Wang et al., 2007) Doozer++ assumes that only positive examples are available for training. The positive-only classification used in PORE may however break in an open IE environment, because it relies on binary classifiers with a strong difference between feature vectors representing positive and negative/unlabeled examples. This difference is likely to diminish as the number of relationship types increases significantly.

Another requirement of our work is to use predictors that allow explicit analysis of the relation-indicative power of specific patterns. For an investigative approach, a white-box architecture based on probabilistic analysis of patterns is more suited than a black-box machine-learning approach such as CRFs in WOE (Wu and Weld, 2010), in which the individual probabilities of specific patterns cannot easily be determined.
5.3 Domain Definition - Hierarchy Creation

This section describes the creation of domain hierarchies as the first step toward domain model creation. Based on the earlier discussion about concept integrity and reference, the approach taken in this work is to extract a domain hierarchy from a larger collection of concepts, namely from the Wikipedia article and category hierarchy. To make this process user friendly and applicable to the task of on-demand creation of domain hierarchies, it was imperative to devise a system that intelligently selects the concepts that are pertinent to a domain solely based on a partial keyword description. It is assumed that a user may only know some of the concepts that are important in a domain, but she still wants a fairly complete, yet focused description of her domain of interest.

There are several methods involved in creating a comprehensive domain hierarchy from a simple set of keywords. The overall process follows an *Expand and Reduce* paradigm that allows us to first explore and exploit the concept space before reducing the concepts that were initially deemed interesting to those that are most important and most significant to the domain of interest.

A domain of interest is looked at from three different levels.

- The Focus, which is at the center of interest.

- The Domain, which encompasses concepts that are immediately related to concepts in the Focus.

- The World View, which prioritizes some concepts and links over others, based on a category that marks the point of view.

The Focus of the domain is described in the form of a keyword query. The Domain is provided in the form of one or more categories that the user wants the model restricted to. The World View is given in the form of a single category that determines the way concepts are categorized. Table 5.1 shows combinations of Focus queries, Domain and WorldView
categories. Since the Focus is described by a query, the first column shows a query syntax. In the Web 2.0 examples, it can easily be seen that the first example aims at creating a model that looks at Web 2.0 from a sociological point of view, whereas the second example gears the creation toward a technical direction.

Table 5.1: Example combinations for Focus, Domain and WorldView

<table>
<thead>
<tr>
<th>Focus</th>
<th>Domain</th>
<th>WorldView</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Web 2.0” AND Netizen</td>
<td>Social Networking</td>
<td>Society</td>
</tr>
<tr>
<td>“Web 2.0” AND Ajax</td>
<td>Internet</td>
<td>Information- Science</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>Oncology</td>
<td>Medicine</td>
</tr>
<tr>
<td>Cognitive</td>
<td>Cognitive Science</td>
<td>Science</td>
</tr>
</tbody>
</table>

Whereas the Focus is based on an individual choice of keywords, Domain and WorldView are dependent on the community-created category hierarchy on Wikipedia and thus emphasize the shared aspect of model creation. The WorldView is generated by topologically sorting the categories of Wikipedia with respect to the chosen category. This sorting is performed by executing a breadth-first search through the category graph starting from the upper category. Only those category links are kept that provide the shortest path from the chosen root category. The assumption behind this uninformed method is that the community that created the topic hierarchy asserted subcategories that are most important to a category closer than subcategories that are only marginally related.

For the steps involved in Expand and Reduce, the ontology creation takes advantage of several well-proven tools for knowledge aggregation:

- **Expansion**
  1. Full-text Search (Banerjee et al., 2007)
  2. Graph-based expansion (Lizorkin et al., 2008),(Turdakov and Velikhov, 2008)
  3. Category-growth

- **Reduction**
1. Category-based reduction/intersection

2. Conditional pruning

3. Depth reduction

In the following, these methods are described in depth, using the following symbols and functions:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C )</td>
<td>Set of concepts without categories</td>
</tr>
<tr>
<td>( C_{\text{search}} )</td>
<td>Set of concepts after Exploration step</td>
</tr>
<tr>
<td>( C_{\text{base}} )</td>
<td>Set of concepts after Exploitation step</td>
</tr>
<tr>
<td>( C_{\text{red}} )</td>
<td>Set of concepts after Expansion and Probabilistic Reduction steps</td>
</tr>
<tr>
<td>( C )</td>
<td>Concept</td>
</tr>
<tr>
<td>( T )</td>
<td>Category hierarchy</td>
</tr>
<tr>
<td>( T_W )</td>
<td>Category hierarchy corresponding to the specified WorldView</td>
</tr>
<tr>
<td>( T_i )</td>
<td>Intermediate category hierarchy</td>
</tr>
<tr>
<td>( M )</td>
<td>Domain Model including individual concepts and categories</td>
</tr>
</tbody>
</table>

5.3.1 Expansion

The expansion steps take the model from a description of a domain in form of a simply keyword query to a comprehensive set of concepts relevant to the domain (Figure 5.3). In the expansion steps recall is maximized to allow as many concepts as possible to be taken into account while maintaining a sensible focus on the domain of interest.

5.3.1.1 Full Text Search – Exploring the knowledge space

For a full-text-search over the content of the Wikipedia articles, the complete set of articles is indexed using the Apache Lucene search engine\(^3\). Any indexed Wikipedia article that matches a query with a score greater than a given threshold \( \epsilon_{\text{search}} \) and/or smaller than

\(^3\text{http://lucene.apache.org/core/}\)
Figure 5.3: Steps 1 (Full-text search) and (Semantic Similarity) in the expansion process

A given maximum rank (depending on user preferences) will be returned, regardless of whether it ultimately matches the desired focus domain or not. However, a carefully stated query will help maintain the focus even in this early stage. The set of concepts \( C_{\text{search}} \) returned from this step is described in Formula 5.1. The user has the option of scored and ranked search because the raw score given by the Lucene indexer is not always intuitive, especially when searching for more elaborate Boolean expressions.

\[
C_{\text{search}}(query) = \{ \text{concept(article), article} \in \text{hits(query)} | \text{score(article)} > \epsilon_{\text{search}} \}
\]

and/or

\[
C_{\text{search}}(query) = \{ \text{concept(article), article} \in \text{hits(query)} | \text{rank(article)} < \text{maxRank} \}
\]

(5.1)

5.3.1.2 Graph-Based Expansion – Exploiting the knowledge space

The graph-based expansion returns articles that are semantically closely related to the initial search results, even if the articles do not match the focus query. The graph-based expansion follows a semantic similarity method, because similarity is computed using links between Wikipedia pages. The importance or similarity of adjacent articles is measured using a weighted common neighbors metric based on (Lizorkin et al., 2008). The semantic
similarity between nodes $a$ and $b$ is given in Equation 5.2, which is defined as the sum of weights of their shared neighbors (articles that are linked to or link to the current article), normalized by the node degrees. Let $M$ be the adjacency matrix of Wikipedia, $N(a)$ stands for the neighborhood and $w(N(a))$ stands for the link weights to $a$’s neighboring nodes, and includes all the articles that link to or are linked to $a$. Thereby, $w(N_i(a))$ describes the link weight to the $i^{th}$ neighbor of $a$. The weights depend on the kind of link between $a$ and each of its neighbors, as shown in Figure 5.4. Equation 5.2 is similar to the first iteration of SimRank (Jeh and Widom, 2002); the difference is in the normalization factor and the weighted links.

The link weights $w$ to the neighboring nodes can vary for different document links considered. Figure 5.4 shows the different types of links in Wikipedia as described in (Turdakov and Velikhov, 2008). The weights for each of these types of links were empirically determined. The scoring emphasizes on the see-also links, because editors add these links usually to refer to highly relevant concepts. Double links also indicate that two concepts are mutually important for each other. By the same rationale single links in one or the other direction are given low weights. Many articles, for example, have a link to years or countries, but only if the linked articles also contain a link back can it be assumed that the original article is actually important for the year, the country, etc. The set $C_{sim}$ of nodes with highest similarity to the initial nodes are then chosen to be in the domain model (see Equation 5.3)
\[ C_{sim} = \{ a \in C_{search}, b \in G_{Wiki} | \text{sim}_{graph}(a, b) > \epsilon_{sim} \} \]  

(5.3)

The final set of concepts gained during the expansion steps is the union of the initial search results and their graph-based expansions (Equation 5.4).

\[ C_{base} = C_{search} \cup C_{sim} \]  

(5.4)

5.3.1.3 Building a category hierarchy

Building a category hierarchy is an essential step for further pruning. Categories are connected with respect to the World View taken, rather than using the entire graph structure of Wikipedia.

Equations 5.5, 5.6 and 5.7 describe the iterative growth of the category hierarchy. In Equation 5.5, the immediate types/categories \( T_1 \) are selected that contain the concepts from \( C_{base} \); Equation 5.6 describes finding the super-categories to the categories in \( T_i \) within the world view \( T_W \) and adding them to the set of categories in the hierarchy. The final category hierarchy \( T \) is reached once \( T_{i+1} = T_i \).
\[ \mathcal{T}_1(C_{base}) = \{T \in \mathcal{T}_W | \exists C \in C_{base} : C \in T \} \] (5.5)

\[ \mathcal{T}_{i+1}(\mathcal{T}_i) = \mathcal{T}_i \cup \{T \in \mathcal{T}_W | \exists T_{sub} \in \mathcal{T}_i : T_{sub} \text{ subCategoryOf } T \} \] (5.6)

\[ \mathcal{T} = \mathcal{T}_{i+1}, \text{ IF } \mathcal{T}_{i+1} \equiv \mathcal{T}_i \] (5.7)

Where \( \mathcal{T}_i \) is the set of categories (i.e. types) at iteration i, \( T \) denotes a single category and \( C \in T \) denotes a concept categorized in category \( T \).

### 5.3.2 Reduction

Whereas the expansion steps are used to gather knowledge in a recall-oriented way, the reduction steps increase precision and reduce the set of concepts to the focus domain.

#### 5.3.2.1 Probability-based reduction: Conditional Pruning

For each concept in the set of extracted concepts \( C_{base} \), two relevance probabilities are computed with respect to the broader domain of interest. Equations 5.8 and 5.9 show these conditional probability computation. Equation 5.8 indicates the importance of a concept for the domain. A probability of 1.0, for example would indicate that every time the concept appears, it is within the domain of interest. Equation 5.9 shows the inverse: how commonly used is the concept in the domain? Knowing both measures is not only important for this pruning step, but also for a potential use of the created domain model in probabilistic document classification tasks. If the importance of a concept is less than one of the predefined thresholds \( \epsilon_1, \epsilon_2 \), it is discarded from the set of domain concepts.

Practically, the relevance probabilities are computed over the Wikipedia link graph \( G_W \). The expression \( \text{link}(a, b) \) in Equations 5.8 and 5.9 means that a link exists from
article \( a \) to article \( b \) on Wikipedia, meaning that the concepts described by \( a \) and \( b \) are related. The denominator in both equations indicates the number of article nodes in \( G_W \) that link to both \( C \) and any concept in \( C_{\text{base}} \).

\[
p(C_{\text{base}}|C) = \frac{|\{a \in G_W \mid \forall d \in C_{\text{base}}, \text{link}(a, C) \land \text{link}(a, d)\}|}{|\{a \in W \mid \text{link}(a, C)\}|} \quad (5.8)
\]

\[
p(C|C_{\text{base}}) = \frac{|\{a \in G_W \mid \forall d \in C_{\text{base}}, \text{link}(a, C) \land \text{link}(a, d)\}|}{|\{a \in W \mid \forall d \in C_{\text{base}}, \text{link}(a, d)\}|} \quad (5.9)
\]

\[
C_{\text{red}}(C_{\text{base}}) = \{C \in C_{\text{base}} \mid p(D|C) \geq \epsilon_1 \land p(C|D) \geq \epsilon_2\} \quad (5.10)
\]

Finally, the domain model \( \mathcal{M} \) is constructed as the union of the set of remaining concepts \( C_{\text{red}} \) and the imposed category hierarchy \( \mathcal{T} \):

\[
\mathcal{M} = C_{\text{red}} \cup \mathcal{T} \quad (5.11)
\]

### 5.3.2.2 Category-based reduction

After probabilistic pruning, some categories (and their subcategories) will be empty. These can by default be deleted. Furthermore, all categories that do not belong to the chosen broader focus domain, including all concepts that are not categorized within the broader focus, are deleted immediately. If a concept is categorized in more than one category, it is kept in the categories that are part of the broader focus domain, otherwise it is deleted. If a category contains only one concept, that concept is moved up to the next higher category.
in the hierarchy and its original category is deleted.

**Depth Reduction**

In many cases, after the category-based reduction, deep linear branches of categories remain as artifacts of the category building and deletion tasks. It is assumed that unpopulated and unbranched category hierarchies can be collapsed without loss of relevant knowledge. This step reduces the depth and increases the fan-out of the domain model. Together with the previous step it reduces the number of resulting categories to make the model easier to manage. Algorithm 1 describes the collapsing of categories.

**Algorithm 1 Depth-Reduction Algorithm**

1: For all categories $T \in M$
2: if $\exists! \text{subCategories}(T) \text{ AND concepts}(T) = \emptyset$ then
3: $T_a = \text{subCategories}(T)$
4: $\forall T_i \in \text{subCategories}(T_a)$ :
5: $\text{subCategories}(T) = \text{subCategories}(T) \cup T_i$
6: remove $T_a$
7: end if

After expansion and reduction, the intermediate model consists of a category hierarchy and concept nodes that are annotated with their relevance probabilities. The concept nodes are named with the corresponding Wikipedia article names. In the following, the nodes are enriched with alternative descriptors.

**5.3.3 Synonym Acquisition**

Since the extraction of domain models is concept-based, rather than term-based, it is necessary to find alternative denotations or labels for the concepts in the hierarchy. The Wikipedia article names are unambiguous identifiers and as such not necessarily of the form we are used to when talking about the concept described by the article. The Wikipedia article for the capital of the United States, for example has the Wikipedia name “Washington,
D.C.” as a unique identifier. We expect to find different identifiers in text, though, such as “Washington” or just “D.C.”. A domain model that is used for information extraction and text classification needs to contain these identifiers and synonyms for the concept of the article. One good source of synonyms is WordNet (Fellbaum, 1998), but it requires to first unambiguously identify a match between a Wikipedia article name and a WordNet Synset, which adds another level of uncertainty. Furthermore, WordNet, despite a great overlap, has a different scope and is more static than Wikipedia. A major reason to extract knowledge from Wikipedia was, however, to have an almost instantaneous update of world-knowledge as new concepts become known or as new events unfold. Anchor texts have proven to be a good indicator for alternative concept labels (Brin and Page, 1998). More than on the web, where the concept described by a page is not entirely certain, on Wikipedia, there is a direct concept associated with the link. Hence we decided to stay within the Wikipedia corpus and analyze the anchor texts that link to the articles that describe the concept. The probability that a term is a synonym of an Article name and hence an identifier of the concept described by the article is given by Equation 5.12, the conditional probability that a term is the anchor text of a link to an article:

\[
p_{syn}(C|L) = \frac{|\text{anchor}(L, C)|}{\sum_{a \in G_w} |\text{anchor}(L, a)|}
\]  

Equation 5.12

In the OWL model of the domain hierarchy, these synonyms are represented as labels as well as individual values of annotation properties that contain both the term and \(p_{syn}\), so the synonyms can be used in extraction (Section 5.4) and other classification tasks.

### 5.3.4 Serialization

The resulting domain models are serialized either as OWL files or as RDF files using SKOS\(^4\) relationships. An OWL serialization reflects a view of the Wikipedia category

\(^4\)http://www.w3.org/TR/skos-reference
graph as a class hierarchy whereas SKOS describes category membership and subcategory relationships that do not follow proper inheritance rules. The serialization based on Semantic Web standards greatly facilitates accessibility, manual modification and visualization. When serializing as an OWL file subcategory relationships are turned into subclass relationships, and thus it is important to be aware of the fact that this may not meet the formal standards of OWL; the Wikipedia category hierarchy is often associative rather than expressing formal is_a relationships. However, knowing about the limitations of the generated models, OWL as the W3C-recommended ontology language is a good way to make the models widely accessible.

The next section will discuss the embellishment of the extracted concept hierarchy with triples using automatically extracted facts involving named relationships.

5.4 Domain Description - Information Extraction

Domain description is the second step of domain model creation. As the name suggests, it is concerned with describing the concepts that were found to be relevant for the domain. This description is achieved by connecting the concepts or individual entities in the domain with named relationships that indicate how one concept or entity relates to another. The relationship types are defined in an implicit or explicit schema, for example in DBpedia (Bizer et al., 2009b) or UMLS (Lindberg et al., 1993). As stated in the overview section, the factual quality of the extraction, apart from the performance of the algorithm, is guaranteed by either extracting from community-created or peer reviewed corpora, such as Wikipedia and MEDLINE\textsuperscript{5}, or by aggregation of evidence for each fact in the case of uncontrolled corpora, such as general Web pages.

A fact is an instantiation of a named relationship, involving a subject and an object concept. In order to be called a fact, it also needs to express a true real-world connection.

\textsuperscript{5}MEDLINE Factsheet: http://www.nlm.nih.gov/ pubs/factsheets/medline.html
An extracted fact is thus a verified extracted statement. Since this work is only concerned with binary relationships, all the statements will formally be represented as \( \langle \text{Subject, Relationship, Object} \rangle \) triples. The triple representation follows the RDF standard that is prevalent in the Semantic Web. Many datasets on the Linked open Data (LoD) cloud are also represented in this format. However, since class membership can be expressed by a binary relationship, this can also be extracted.

Fact extraction is implemented as a classification of concept pairs into relationship types. Since each concept pair can potentially participate in multiple relationships (e.g. \textit{birthplace} and \textit{residence}), a probabilistic multi-class relationship classifier was developed that operates on a sparse concept-pair to pattern vector space representation. One impediment that has troubled purely pattern-based approaches to IE is low recall. This was improved by using a pattern-generalization step. Another challenge faced was that of classifying into multiple classes of relationships where appropriate. This was addressed using a statistical pertinence measure that reevaluates the importance of a pattern to a relationship class based on the similarity to other classes.

Using only positive examples of features and having only positive examples of relationship types, the classifier must face the problem of missing discriminating data. Still, with the large and ever growing LoD datasets at hand, the number of relationship types that the pattern occurrences are classified into also grows and thus our ability to discriminate between positive examples. The challenge is to discriminate relationships that seem very similar to the classifier because of the above mentioned problems. The main goal of a positive-only classifier is thus emphasizing differences and penalizing similarities between different relationships, while not penalizing similarities between similar relationships. One unique feature of this classifier is thus an increase in discriminative performance as the number of relationships that are classified grows. This is a significant improvement over traditional supervised IE techniques that tend to have better performance when only a limited number of classes are trained on.
This section is structured as follows. First, it will introduce the notion of surface patterns. Then it will give a general idea of the probabilistic framework, before showing how this is represented in a vector space. Subsequently it will explain the pertinence measure and show examples how this measure impacts the importance of individual patterns. Finally it will show how the classifier is built and applied.

5.4.1 Surface Patterns

The pattern-based IE algorithm developed for this work uses distant supervision (Mintz et al., 2009) to extract patterns that express formal relationships in free text. When a phrase is found in text that potentially expresses a fact from the training corpus, it is generalized into a pattern by replacing the terms that indicate the subject/object pair of the fact with placeholders. Hence, a pattern is of the form “[Prefix]⟨L1⟩[Infix]⟨L2⟩[Postfix]” with ⟨L1, L2⟩ indicating a pair of concept labels found using Equation 5.12 that denote subject and object concepts of the triple. The textual grounding of a triple is the collection of phrases that express the semantic content of the formal statement.

As an example consider the triple ⟨Albert Einstein, birthPlace, Ulm⟩ and potential textual manifestations of this fact. The phrase “Albert Einstein was born in Ulm” is added to the pattern dictionary as \(P_i\): “⟨Subject⟩ was born in ⟨Object⟩”. Similarly, the phrase “Ulm is the birthplace of Albert Einstein” is represented as \(P_j\): “⟨Object⟩ is the birthplace of ⟨Subject⟩”. Assuming that these patterns occur in the frequencies \(f_i\) and \(f_j\) respectively, the triple ⟨Albert Einstein, birthPlace, Ulm⟩ is then grounded by the textual occurrences \(⟨P_i, f_i⟩\) and \(⟨P_j, f_j⟩\).

The top-down step to domain definition gives us the probabilities with which \(L_1\) and \(L_2\) denote the proper concepts mentioned in the triple, because concept labels and the probability that they actually refer to the concept are extracted along with the concepts. Section 5.4.2 will show how these probabilities are used.
5.4.1.1 Pattern Generalization

A major drawback of surface patterns is their lack of variability and associated low recall. Apart from such widely applicable patterns as “⟨Subject⟩ was born in ⟨Object⟩”, there will be many very specific patterns that are only applicable to few concept pairs, such as “⟨Subject⟩ graduated in 1998 from ⟨Object⟩”. An ideal generalization of this pattern to indicate an almaMater relationship is “⟨Subject⟩ graduated in * from ⟨Object⟩”. This is arguably the best generalization of the original pattern. However, pattern extraction is agnostic to the semantics or even the part of speech or sentence position of the generalized tokens and hence the generalization is also unaware of to the semantics of the tokens that are generalized. Since no parse tree exists, the kinds of generalization that are used in NLP-based approaches, e.g. (Suchanek et al., 2006), where shortest paths through parse trees are used to generalize patterns can also not be used as a guide. For this reason, all possible generalizations have to be created and then evaluated with respect to their predictive power for a relationship. Patterns with low predictive power will be pruned. After pruning, the set of remaining generalized patterns approximates the kind of generalization that is done in NLP-based generalization.

Table 5.3 shows an example of a complete generalization, where an exponential $2^{\lvert tokens \rvert}$ patterns are produced. To save space an example pattern with a 4-token infix pattern was chosen, leaving out prefix and postfix. In practice, most of the patterns that contain more than 2 wildcards are pruned during minimization, because they are insignificant indicators for specific relationships. Thus the application usually limits the generalization to $\lfloor \lvert tokens \rvert / 2 \rfloor$ wildcards per pattern, resulting in $\hat{n}$ patterns. In Table 5.3, this corresponds to rows 1-7, 9-11 and 13. Equation 5.13 computes the number of generalized patterns that are created from a raw pattern, when the number of wildcards is restricted, leading to a slower growth in the number of patterns.

$$\hat{n} = \left(\left\lfloor \frac{\lvert tokens \rvert + \lvert wildcards \rvert}{\lvert wildcards \rvert} \right\rfloor - 1\right)$$  \hspace{1cm} (5.13)

115
Table 5.3: Pattern-generalization example

<table>
<thead>
<tr>
<th></th>
<th>Subject</th>
<th>graduated</th>
<th>in 1998</th>
<th>from</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>⟨</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>⟨</td>
<td></td>
<td></td>
<td>*</td>
<td>⟨</td>
</tr>
<tr>
<td>3</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>from</td>
<td>⟨</td>
</tr>
<tr>
<td>4</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>⟨</td>
</tr>
<tr>
<td>5</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>1998</td>
<td>from</td>
</tr>
<tr>
<td>6</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>1998</td>
<td>*</td>
</tr>
<tr>
<td>7</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>from</td>
</tr>
<tr>
<td>8</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>9</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>in</td>
<td>1998</td>
</tr>
<tr>
<td>10</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>in</td>
<td>1998</td>
</tr>
<tr>
<td>11</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>in</td>
<td>*</td>
</tr>
<tr>
<td>12</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>in</td>
<td>*</td>
</tr>
<tr>
<td>13</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>1998</td>
</tr>
<tr>
<td>14</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>1998</td>
</tr>
<tr>
<td>15</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>16</td>
<td>⟨</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

5.4.2 Probabilistic Framework

A probabilistic classifier for relationships intuitively answers the question: “Which relationship is likely expressed when two entities appear with these patterns?” It is also manually verifiable, which makes it a good candidate for a prototype application.

Before discussing solutions to this problem the general idea of the probabilistic classification is outlined.

Table 5.4: Terminology

<table>
<thead>
<tr>
<th>C</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Set of all concepts</td>
</tr>
<tr>
<td>S, O</td>
<td>Subject/Object concept of the triple</td>
</tr>
<tr>
<td>L</td>
<td>Term/Label</td>
</tr>
<tr>
<td>L_S, L_O</td>
<td>Term expressing the subject/object of a triple</td>
</tr>
<tr>
<td>T_S, T_O</td>
<td>Semantic type/class of the subject/object</td>
</tr>
<tr>
<td>P</td>
<td>Surface pattern</td>
</tr>
<tr>
<td>R</td>
<td>Relationship type</td>
</tr>
<tr>
<td>R</td>
<td>Set of all relationship types</td>
</tr>
<tr>
<td>R2P</td>
<td>Relationship-Pattern matrix</td>
</tr>
<tr>
<td>CP2P</td>
<td>Concept-Pair-Pattern matrix</td>
</tr>
<tr>
<td>CP2R</td>
<td>Concept-Pair-Relationship matrix</td>
</tr>
<tr>
<td>M_D, M_R</td>
<td>Domain and Range prior probability matrices</td>
</tr>
</tbody>
</table>

Derivation of the Classifier The following paragraphs will show the derivation of the probabilistic extraction framework. IE is hereby cast as a classification from concept
pairs into relationship types using surface patterns as features. The terminology used in the derivation is given in Table 5.4.

It is important to observe that a distinction is made between the concepts that participate in the relationship and the labels that denote the concepts in text. Similarly, the extraction algorithm operates on concepts rather than labels. However, since concepts do not actually appear in text, the grounding of the concepts in text has to be done using the denoting labels. Given the ambiguity in mapping, a single occurrence of a term pair in text is not enough to indicate the concept pair that is sought. Similarly to the mapping of pattern occurrences to types of relationships, the mapping of multiple occurrences different terms to one concept assure the proper references.

Suppose a given concept pair \( \langle S, O \rangle \) from the set of all concepts on Wikipedia \( C \) that should be classified into one or more relationship types \( R \). The general goal is hence to find a solution to the conditional probability \( p(R|S,O) \) and identify all relationships \( R_{S,O} \) that have a sufficiently high confidence of relating \( S \) to \( O \) (Equation 5.14).

\[
R_{S,O} = \{ R \in R | p(R|S,O) > \epsilon_{rel} \}
\]  

(5.14)

The features of the classifier are the patterns found in text that are potential manifestations of the concept pair and the relationship as described in section 5.4.1. Free text, however, will contain ambiguous terms that denote \( S \) and \( O \). Section 5.3.3, explained how to find probabilities for synonyms of article names on Wikipedia in Equation 5.12.

Ideally, the joint probability that a term pair indicates a concept pair would be computed, i.e. \( p(S,O|L_S, L_O) \), because both terms help at mutually disambiguating each other. For example, a term pair \( \langle table, chair \rangle \) puts both terms in the realm of furniture, whereas the term pair \( \langle table, column \rangle \) puts them in the spreadsheet or database field. Unfortunately this joint probability is a) expensive to obtain, considering that over 9 million terms were
identified that link to about 3.7 million Wikipedia articles b) restricts the classifier to term pairs that have already been analyzed and c) is even more corpus-dependent than finding the probability that a single term denotes a single concept. For this reason we assume conditional independence of $p(S|L_S)$ and $p(O|L_O)$.

A factor that helps in classification, but is dependent on the availability of background knowledge, is the information on domain and range of relationships, expressed here as the probability $p(R|T_S,T_O)$ of seeing a relationship given a domain $T_S$ and a range $T_O$. Here, as well, the joint probabilities are too restrictive for an open extraction so we assume independence of $p(R|T_S)$ and $p(R|T_O)$.

Depending on the background knowledge that is present in the form of ontologies, taxonomies and dictionaries, the terms $p(S|L_S)$, $p(O|L_O)$, $p(R|T_S)$ and $p(R|T_O)$ may be only partially available or not at all. In this case the classifier operates purely on the language model, rather than on the combined semantic model/language model. In a well-designed ontology, relationships will be assigned domains and ranges. However, in the case of community-created sources such as DBpedia this is more difficult. Even though an ontology exists that covers the entities on DBpedia, it is too coarse-grained to properly match the models produced in the Domain Definition step and also does not assign domain- and range restrictions to all properties. In this case, the probabilities $p(R|T_S)$ and $p(R|T_O)$ can be derived bottom-up, by analyzing the category-coverage of the facts in the KB as shown in Section 5.4.5.

The third and central component of the probabilistic framework are the relationship-pattern probabilities $p(R|P)$, i.e. the probability of seeing a relationship in the presence of a specific pattern or a vector of relationship probabilities given a vector of pattern frequencies. Separating $p(R|P)$ allows us to build a fixed pattern representation for relationships.

Figure 5.5 depicts a Bayesian Network that graphically models the classifier, showing how it operates on a Semantic model and a statistical language model in a unified manner. The probability $p(R,S,O)$ of a relationship occurring with a subject and an object can
be rewritten as \( p(R, S, O, P, L_S, L_O, T_S, T_O) \), which, based on the Bayesian Network, is formalized in Equation 5.15. The equation makes use of the independence assumption of \( p(S|L_S), p(O|L_O), p(R|T_S), \) and \( p(R|T_O) \) and approximates \( p(R|T_S, P, T_O) \) as the product of \( p(R|T_S), p(R|T_O) \) and \( p(R|P) \). Specifically, as the probability of the presence of a relationship \( R_j \) between \( S \) and \( O \) is computed over all the patterns that \( L_S \) and \( L_O \) appear in, the classifier sums over the probabilities of all occurrences of a pattern with \( S \) and \( O \), each weighted by the probability that its pattern indicates the relationship \( R \). For the types, the probability is maximized over the domain and range types, indicating that the \( T_S \) and \( T_O \) form hierarchies and the type that has the strongest support as a domain or range for \( R \) is chosen.

\[
p(R, S, O) \approx \sum_{L_S \in S} \sum_{L_O \in O} \sum_{P \in \text{docs}} p(P|L_S, L_O) \cdot p(R|P) \cdot p(S|L_S) \cdot p(O|L_O) \cdot \max_{t_S \in T_S} p(R|t_S) \cdot \max_{t_O \in T_O} p(R|t_O)
\]  

(5.15)
The values for \( p(R|P) \) are the most difficult to derive, because a distant supervision approach is used for training without access to negative training data to learn these probabilities. Also, as mentioned above, no apriori knowledge is assumed of the relationship semantics and thus the extent of the semantic overlap between relationships needs to be obtained during training. In the following I will describe the distantly supervised training process to create a vector space representation of \( p(R|P) \). The next subsection describes the general acquisition procedure and subsection 5.4.4 details the derivation of the pertinence measure that modifies \( p(R|P) \) to account for intensional and extensional relational similarity.

5.4.3 Vector-Space model

To get a vector-space representation of the probabilistic model that describes \( p(R|P) \), a distantly supervised training procedure is first used to accumulate patterns for individual fact occurrences in a \( \langle \text{Concept-Pair, Pattern frequency} \rangle \) matrix \( CP2P \). The row vectors represent the frequencies of the patterns in which a concept pair appears. During training these are the \( \langle S, O \rangle \) pairs found in LoD triples. The frequencies are accumulated into a \( \langle \text{Relationship, Pattern} \rangle \) matrix \( R2P \) that can be seen as a language model for relationship mentions in text. In the application phase patterns between previously unseen concept pairs are compared to \( R2P \) to yield candidate relationship types the concept pair participates in.

The following steps detail the matrix creation:

1. Find pattern representations of training facts in the text corpus as described in Section 5.4.1 using distant supervision (Mintz et al., 2009). For every textual manifestation of a fact, replace the terms denoting subject and object in the triple with \( \langle \text{Subject} \rangle \) and \( \langle \text{Object} \rangle \) placeholders to generate a pattern. If it is a new pattern, it is added to the pattern dictionary. The internal representation of a fact then becomes a vector that maps the Subject-Object concept pair to a vector of pattern frequencies. A Concept pair to pattern matrix \( CP2P^R \) combines the individual vectors (Equation 5.16).
cates that in the training phase the vectors contain information about the relationship in the triple, so accumulated representations for relationships can be derived. The same procedure is however used during application, when only the concept pairs are known and the relationship needs to be extracted. The weighted frequency for the occurrence of a concept pair \( \langle S, O \rangle_i \) and a pattern \( P_j \) is defined as the product of the probability that a term-set pair \( \{L_S|L_S \text{ label of } S\} :: \{L_O|L_O \text{ label of } O\} \) indicates \( \langle S, O \rangle_i \) and the frequency of seeing the pattern with any of these term pairs.

\[
CP2P_{ij}^R = \text{weighted frequency}(\langle S, O \rangle_i, P_j) =
\sum_{\langle L_S, L_O \rangle \in \langle S_i, O_i \rangle} (|P_j^{L_S, L_O}| \cdot p(S|L_S) \cdot p(O|L_O)) \tag{5.16}
\]

(2) Generalization - The coverage of patterns is increased by substituting tokens in the pattern with wildcard characters as described in Section 5.4.1.1. Usually a generalized pattern is derived from multiple original patterns. The frequencies of the generalized pattern is then computed by adding the frequencies of these original patterns. For example, if the frequency for “\langle Subject \rangle graduated in 1998 from \langle Object \rangle” is 5 and the frequency for “\langle Subject \rangle graduated in 2000 from \langle Object \rangle” is 7, then the frequency of the generalized pattern “\langle Subject \rangle graduated in * from \langle Object \rangle” is 12.

(3) To construct a matrix that contains the probabilities \( p(R|P) \) of seeing a relationship type when encountering a pattern, we first build a matrix that contains the probabilities \( p(P|R) \). This is done by adding all concept pair vectors in \( CP2P_{ij}^R \) that indicate one relationship type into one row vector in \( R2P \) and then normalizing the row vectors. According to Equation 5.17, all vectors in \( CP2P_{ij}^R \) that are annotated with the \( k^{th} \) relationship are added to row \( R2P_k \).
Additionally, the following frequency criteria need to be met:

- for each pattern, the overall weighted frequency is above a threshold $t_1$
- the number of distinct subject-object pairs that the pattern occurs with is above a threshold $t_2$

Taking both the raw frequency of the pattern and the frequency of the training facts that lead to the pattern into account reduces noise by ensuring that a pattern not only occurs often enough, but is also not specific to a particular concept pair.

The algorithm assumes a uniform distribution of relationship types in the world. Even though there is a nonuniform distribution of these types in a fact corpus, it is likely that this distribution is skewed. In the case of Wikipedia/DBpedia, for example, a majority of facts stem from sports-related descriptions. However, this distribution is skewed. With a potentially unbounded number of relationship types in reality, it is safe to assume that the prior probability for each individual type will eventually approach 0. Thus, the more relationship types are available for classification, the more realistic a uniform distribution becomes. For this reason the rows in the $R2P$ matrix are normalized (Equation 5.18), before computing each field $a_{ij}$ in $R2P$ as the probability of seeing $R_i$ when encountering $P_j$ in text (Equation 5.19).

\[
R2P_{freq}^{k} = \sum_{i=1}^{n} CP_i^{rel_k} \tag{5.17}
\]

\[
R2P_{ij}^{norm} = p(P_j|R_i) = \frac{R2P_{ij}^{freq}}{\sum_{k=1}^{n} R2P_{ik}^{freq}} \tag{5.18}
\]

\[
R2P_{ij} = p(R_i|P_j) = \frac{p(P_j|R_i)}{\sum_{k=1}^{m} p(P_j|R_k)} \tag{5.19}
\]
Minimization: prune low probability patterns to reduce noise and to reduce the size of the matrix. The minimum frequency requirements in step (3) ensured that infrequent patterns that were too specific are not taken into consideration. Here, patterns that are too general to be of value for the classification are removed.

The resulting matrix $R2P$ that maps relationships to their occurrence with specific patterns is basically a pattern-based language model designed to detect occurrences of relationships. It is of fairly low complexity, because instantiated patterns can be seen as 3-grams, consisting of subject designator, object designator and relationship designator.

Table 5.5: Extensional and Intensional similarity between relationships. The top half shows taxonomy relationships specific to the biology domain, the lower half shows domain-independent relationships. The first 2 columns show the relationship pair, the next 4 columns show extensional attributes with the number of shared subject-object pairs and the overall number of instantiating facts for each relationship type. Fraction$_{min}$ indicates the fraction of overlap measured by the relationship with the least number of instances. The next 3 columns indicate intensional similarity, computed using pertinence ($+sim_{int}$), omitting pertinence ($-sim_{int}$) and the difference in similarity. A positive value indicates that $+sim_{int}$ assigned higher similarity than $-sim_{int}$ and vice versa. The last column indicates $sim_{rel}$, the overall relational similarity, taking extensional and intensional similarity into account.

<table>
<thead>
<tr>
<th>Relationship Pair</th>
<th># Shared</th>
<th># Rel-1</th>
<th># Rel-2</th>
<th>Fraction$_{min}$</th>
<th>$+sim_{int}$</th>
<th>$-sim_{int}$</th>
<th>Diff</th>
<th>$sim_{rel}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>family genus</td>
<td>603</td>
<td>310775</td>
<td>110358</td>
<td>0.005464</td>
<td>0.10496</td>
<td>0.09192</td>
<td>0.01304</td>
<td>0.10438</td>
</tr>
<tr>
<td>order class</td>
<td>391</td>
<td>184907</td>
<td>179328</td>
<td>0.002180</td>
<td>0.05780</td>
<td>0.04194</td>
<td>0.01586</td>
<td>0.05768</td>
</tr>
<tr>
<td>order family</td>
<td>812</td>
<td>184907</td>
<td>310775</td>
<td>0.004391</td>
<td>0.05089</td>
<td>0.03998</td>
<td>0.01090</td>
<td>0.05067</td>
</tr>
<tr>
<td>kingdom class</td>
<td>1</td>
<td>168222</td>
<td>179328</td>
<td>0.000006</td>
<td>0.03572</td>
<td>0.03784</td>
<td>-0.00211</td>
<td>0.03572</td>
</tr>
<tr>
<td>family class</td>
<td>228</td>
<td>310775</td>
<td>184907</td>
<td>0.001271</td>
<td>0.03114</td>
<td>0.03177</td>
<td>-0.00062</td>
<td>0.03110</td>
</tr>
<tr>
<td>phylum order</td>
<td>51</td>
<td>129784</td>
<td>184907</td>
<td>0.000393</td>
<td>0.00873</td>
<td>0.01234</td>
<td>-0.00360</td>
<td>0.00873</td>
</tr>
<tr>
<td>producer artist</td>
<td>21120</td>
<td>160816</td>
<td>99110</td>
<td>0.213097</td>
<td>0.36109</td>
<td>0.33419</td>
<td>0.02689</td>
<td>0.28414</td>
</tr>
<tr>
<td>formerTeam team</td>
<td>21466</td>
<td>68742</td>
<td>337707</td>
<td>0.312269</td>
<td>0.34508</td>
<td>0.31623</td>
<td>0.02884</td>
<td>0.23732</td>
</tr>
<tr>
<td>musicalArtist writer</td>
<td>8628</td>
<td>36606</td>
<td>111408</td>
<td>0.235699</td>
<td>0.33847</td>
<td>0.33306</td>
<td>0.00541</td>
<td>0.25869</td>
</tr>
<tr>
<td>nationality birthPlace</td>
<td>9340</td>
<td>44606</td>
<td>432741</td>
<td>0.209389</td>
<td>0.09764</td>
<td>0.10227</td>
<td>-0.00462</td>
<td>0.07720</td>
</tr>
<tr>
<td>writer director</td>
<td>17968</td>
<td>111408</td>
<td>60210</td>
<td>0.29422</td>
<td>0.08033</td>
<td>0.06330</td>
<td>0.01702</td>
<td>0.05636</td>
</tr>
<tr>
<td>developer publisher</td>
<td>4936</td>
<td>21467</td>
<td>36909</td>
<td>0.229934</td>
<td>0.06114</td>
<td>0.06794</td>
<td>-0.00680</td>
<td>0.04708</td>
</tr>
<tr>
<td>birthPlace deathPlace</td>
<td>25485</td>
<td>432741</td>
<td>104257</td>
<td>0.244444</td>
<td>0.05543</td>
<td>0.06178</td>
<td>-0.00634</td>
<td>0.04188</td>
</tr>
</tbody>
</table>

5.4.4 Pertinence

As described thus far, the distantly supervised training process implicitly made the naïve assumption that an instance of a relationship is determined by subject and object alone. This assumption obviously does not always hold. For example, the fact corpus may contain a
statement about a person being born in one place and another statement about the person having died in the same place. Since the pattern-finding algorithm is agnostic about the semantics of a pattern, it will extract the same patterns for both facts. An analysis of the preprocessed DBpedia Infobox facts shows that out of 3,544,160 facts in the corpus there are 846,574 subject-object pairs that occur with more than one relationship. Overlap between relationships by virtue of their subject-object pairs is referred to as extensional similarity. The lower half of Table 5.5 shows some examples for extensionally similar relationships. For example, there is about 30% overlap between the writer and director relationships as well as the formerTeam and team relationships.

An intuitive solution to avoiding errors due to extensional overlap is to leave out duplicate subject object pairs from the training data. However, it is likely that multiple relationships and their textual representations still exist between many of the remaining concept pairs even when they are not formalized in the training set. A more robust solution is to include duplicates in the training and have measures that can detect both kinds of similarity after the pattern extraction. This is especially important since LoD is seen as streaming, rather than static data and it is assumed that at runtime no informed decision can be made about which facts should be considered for training and which should not. Since LoD only contains positive assertions, the algorithm also cannot rely on negative examples to resolve the ambiguities that inevitably will occur in the extracted patterns. Here, the challenge is thus to identify and weaken the predictive power of patterns that were incorrectly extracted for a relationship type, because they belong to another relationship that just happened to be extensionally overlapping.

In addition to extensionally similar types of relationships, many others are intensionally similar. For example, physical part of entails part of. Intensional similarity gives a different challenge to a classifier, because here the classifier should not just emphasize the patterns that distinguish the relationships, but also maintain predictive power of patterns that indicate both relationships.
The pertinence measure for relationship patterns was developed to account for semantically similar relationship types. It is inspired by the pertinence measure for term-pairs described in (Turney, 2006). It boosts the probability of patterns, if they have a high probability of indicating a specific relationship, even though the pattern is shared among different relationship types. Conversely, it diminishes the probability of patterns that are shared because of extensional overlap.

The intuition behind pertinence is that of probability over a semantic space rather than a fixed class. Usually, a probabilistic approach assigns probabilities to each feature (in this case a pattern) in such a way that the probabilities for each class the feature participates in add up to 1. Using Pertinence, these values can add up to more than 1 when the relationship classes it indicates are semantically overlapping. Pertinence has the effect that similar relationships do not penalize each others shared patterns, whereas dissimilar relationships that share the same patterns get lower scores for these patterns.

The pertinence measure in Equation 5.20 achieves this adjustment of pattern probabilities. The measure is technically a modified conditional probability computation using a weighted sum in the denominator to take relationship similarities into account.

\[
\tilde{p}(R_i | P_j) = \frac{p(P_j | R_i)}{\sum_{k=1}^{m} p(P_j | R_k) \cdot g(1 - \text{sim}_{rel}(R_k, R_i))}
\]

(5.20)

The factor \((1 - \text{sim}_{rel}(R_k, R_i))\) grows the more dissimilar a relationships \(R_k\) is to \(R_i\) and thus reduces the impact of \(P_j\) on \(R_i\). The function \(g : [0..1] \rightarrow [0..1]\) can be any monotonous weighting function. It has proven useful to use a logistic function that amplifies closeness and distance of vectors.

Intensional similarity yields a similar textual representation of relationships in the form of patterns (Turney, 2006). Therefore, the \(\text{sim}_{rel}\) function can compute similarity based on the pattern-probability-vector representation of the relationships. Turney (Turney, 2006) uses a cosine distance between row vectors that represent analogous word pairs. Here, the cosine between rows that represent relationships is computed on the Singular
Value Decomposition (SVD) of the $R2P$ matrix, because SVD inherently identifies highly
descriptive latent dimensions in the data, is known to detect co-importance of features and
helps to reduce noise. Since the SVD matrix does not follow the probabilistic framework
it is only used in this similarity computation, instead of performing the classification itself
using a decomposed matrix. The intensional similarity ($\text{sim}_{\text{int}}$) computation is shown in
Equation 5.21, where $\cos_{\text{SVD}}(R_k, R_i)$ is the cosine of the pattern vectors that describe the
relationships $R_k$ and $R_i$ in the SVD decomposition of $R2P$ (described in section 5.4.3).

$$\text{sim}_{\text{int}}(R_k, R_i) = \cos_{\text{SVD}}(R_k, R_i)$$ (5.21)

The inverse to the premise that intensionally similar relationships share the same pat-
terns does not hold, as purely extensional similarity between two relationships (e.g. the
birthPlace and deathPlace relationships) also yields shared patterns. Thus a counterbal-
ance has to be found to the $\text{sim}_{\text{int}}$ measure. Extensional similarity can be computed by
dividing the number of shared $\langle S, O \rangle$ pairs by a function $f$ of the total number of facts
that instantiate the two relationships (Equation 5.22). The impact of the choice of $f$ can be
compared in Table 5.5. The “Fraction$_{\text{min}}$” column shows the $\text{sim}_{\text{ext}}$ score for $f(R_k, R_i)\equiv min(|\hat{R}_k|, |\hat{R}_i|)$. Other options for this function are $f(R_k, R_i) = \text{avg}(|\hat{R}_k|, |\hat{R}_i|)$ and
$f(R_k, R_i) = \text{max}(|\hat{R}_k|, |\hat{R}_i|)$, where $\hat{R}$ indicates the extension of $R$.

$$\text{sim}_{\text{ext}}(R_k, R_i) = \frac{|\{r_k(a,b) \in \hat{R}_k | \exists r_i(a',b') \in \hat{R}_i, a = a', b = b'\}|}{f(R_k, R_i)}$$ (5.22)

The overall relational similarity is then computed as the intensional similarity $\text{sim}_{\text{int}}$
weighted by 1 minus the extensional similarity $\text{sim}_{\text{ext}}$. See Equation 5.23. The weight-
ing of the intensional similarity diminishes a false attribution of relational similarity, just
because $\langle S, O \rangle$ pairs are shared.
\[
sim_{rel}(R_k, R_i) = \sim_{int}(R_k, R_i) \cdot (1 - f_{ext}(\sim_{ext}(R_k, R_i))) \tag{5.23}
\]

The function \(f_{ext} : [0..1] \rightarrow [0..1]\) is used to adjust the importance of the extensional similarity weight. A logistic function is a good choice here, as well. Analogous to the computation of the \(R2P\) matrix in Equation 5.19, the pertinence-adjusted matrix \(\hat{R2P}\) is computed according to Equation 5.24 using the pertinence computation from Equation 5.20:

\[
\hat{R2P}_{ij} = \hat{p}(R_i|P_j) \tag{5.24}
\]

### 5.4.4.1 Pertinence Analysis

Table 5.5, besides the above described extensional measures, shows the intensional similarity computed by the cosine distance between relationship vectors. The three rightmost columns in the table show the similarity after applying pertinence \((+\sim_{int})\), before applying pertinence \((-\sim_{int})\) and the difference in similarity. A positive difference indicates that the pertinence computation yielded a greater similarity than non-pertinence and vice versa.

The top half of Table 5.5 shows examples of taxonomic relationships from the biology domain. The pertinence computation adjusted pattern probabilities such that the top three relationships pairs became more similar. These three pairs are also immediate ancestors in the taxonomic classification (Species ⇒ Genus ⇒ Family ⇒ Order ⇒ Class ⇒ Phylum ⇒ Kingdom ⇒ Domain), whereas the others are at least one step removed. The lower half of Table 5.5 gives a good indication of the ameliorating effect of pertinence on relationships that have high extensional overlap. All \(\sim_{rel}\) values are significantly lower than the original \(-\sim_{int}\) values.

Figure 5.6 shows the impact of pertinence on fact extraction. As anticipated, per-
Figure 5.6: Comparing precision and recall of fact extraction with and without pertinence. Pertinence has most influence in high-recall regions. Intuitively, as the confidence threshold is increased, patterns that are highly indicative of specific relationships contribute more to the classification and thus the impact of pertinence is diminished.

Pertinence leads to higher recall in the lower-confidence regions, because the probability of individual features is increased. However, because of the extensional similarity measure that is applied as part of the pertinence computation, some pattern-probabilities are lower, which leads to higher precision throughout the range of confidence thresholds.

5.4.5 Relationship Domain and Range probabilities

Often the type of relationship we want to express in natural language is not given by a pattern in a sentence alone, but is also dependent on context. We know that the verb “broke” indicates a different predicate in “Marcus broke the glass” and “Marcus broke the law”. The context here is that the law can be subject to violation and a glass can be shattered. This context is part of our world knowledge, but it can also formally be expressed in ontologies. A classifier needs to know how likely it is that glass or law is an object to the relationship violate or the relationship shattered to guide classification in the right direction.

In well-defined ontologies, domain and range of relationship types are provided in the form of restrictions. A relationship birthplace, for example is defined for the domain Per-
son and the range Location. For LoD data a well-defined ontology may not be present and the possibility of seeing a relationship between instances of two classes must be inferred. Equation 5.25 shows the conditional probability computation of a relationship $R$ given the types $T_S$ and $T_O$. Triples are shown as $\langle S, R, O \rangle$, with an asterisk indicating a wildcard that fits all possible subjects, objects or relationships. $S \in T_S$ and $O \in T_O$ indicate that $T_S$ and $T_O$ are direct or indirect types of a subject $S$ or object $O$. $G$ indicates all triples in the training data.

\[
P_{\text{domain}}(R|T_S) = \frac{|\{\langle S, *, * \rangle \in G | S \in T_S \}|}{|\{\langle *, *, * \rangle \in G | S \in T_S \}|}
\]

\[
P_{\text{range}}(R|T_O) = \frac{|\{\langle *, R, O \rangle \in G | O \in T_O \}|}{|\{\langle *, *, O \rangle \in G | O \in T_O \}|}
\]

Even though an ontology exists for the DBpedia dataset, it is not yet comprehensive enough to account for all entities in DBpedia and it does not provide restrictions for all types of relationships. Moreover, and most importantly, the classes in the ontology do not fully reflect the category hierarchy on Wikipedia that is used for the domain hierarchy creation in Section 5.3. The domain-range probability computation assures that an estimate for all types of relationships that can be encountered at every category on Wikipedia can be found. The challenge in this computation is the nature of the Wikipedia category graph, which is highly interconnected and does not provide a tree or even lattice structure. Whereas it is tempting to think of the category hierarchy as a class hierarchy, the category links are often rather associative in nature and do not express type or inheritance relationships. For example, Sir Tim Berners-Lee is categorized, amongst others in both “British Computer Scientists”, which is a correct classification, as well as “HTTP”, which is an associative relationship. Taking the latter as a classification makes Tim Berners-Lee a type of Internet protocol and conversely assigns a possibility of an Internet protocol having a birth place
or spouse. However, despite this kind of noise, the domain and range probabilities are higher for correct domain and range classes. This is sufficient for the domain and range probabilities because they are merely meant to steer the classification in the right direction when the patterns alone are ambiguous.

In the vector space representation, the values for $p(R|T_S)$ and $p(R|T_O)$ are stored in 2 relationship-prior matrices. The relationship-domain-prior matrix $M_D$ is then of the form $M_{D_{ij}} = \max_{t \in T_{S_i}} p(R_j|t)$ and the relationship-range-prior matrix $M_R$ is of the form $M_{R_{ij}} = \max_{t \in T_{O_i}} p(R_j|t)$. The $p(R|T_S)$ and $p(R|T_O)$ values can be pre-computed for all domain and range classes over all relationships from the LoD data. The specific $M_D$ and $M_R$ matrices are then filled with these values.

### 5.4.6 Matrix-Based Fact Extraction

The extraction of new statements from text is formalized as a classification of concept pairs into relationship types. Given a set of concept pairs, pattern-features for these pairs are extracted from text as described in Section 5.4.3, step (1). This gives us a concept pair-pattern matrix $CP2P$ that contains weighted pattern frequencies. In order to have a row sum of 1, the rows in the matrix are normalized according to Equation 5.26 such that every field in $\tilde{CP2P}$ contains the probabilities $p(P_j|(S,O)_i)$. Equation 5.27 then computes the probabilities of each relationship being instantiated by the concept pair using the pertinence computation from Equation 5.20 for $\tilde{p}(R_j|P_k)$.

$$\tilde{CP2P}_{ij} = p(P_j|(S,O)_i) = \frac{CP2P_{ij}}{\sum_{k=1}^{n} CP2P_{ik}} \quad (5.26)$$

$$p(R_j|(S,O)_i) = \sum_{k=1}^{m} p(P_k|(S,O)_i) \cdot \tilde{p}(R_j|P_k) \quad (5.27)$$

In practice, Equation 5.27 is computed using matrix multiplication (Equation 5.28).

Hence the probabilities of every concept pair instantiating each relationship type is done
in one computation step. The resulting concept pair - relationship matrix can thus also be expressed in terms of Equation 5.27, i.e. \( CP2R_{ij} = p(R_j|(S,O)_i) \).

\[
CP2R = CP2P \times \tilde{R2P^T} \tag{5.28}
\]

Taking the domain and range probabilities into account, the final matrix \( \tilde{CP2R} \) is then computed by performing a Hadamard (entry wise) multiplication over \( CP2R \) and the domain and range prior matrices \( M_D \) and \( M_R \) (Equation 5.29).

\[
\tilde{CP2R} = CP2R \circ M_D \circ M_R \tag{5.29}
\]

The complexity of computing \( CP2R \) is \( O(nmp) \) with \( n \) being the number of concept pairs, \( m \) the number of relationships and \( p \) the number of patterns. However, since the pattern representations for each concept pair are very sparse, the average complexity is of the order \( O(nmc) \) with \( c \ll p \). We found that \( c \) is on average 11.4 based on the distinct patterns found per concept pair. That the complexity is in practice square rather than cubic makes the algorithm very efficient.

### 5.4.7 Pattern Analysis

The probabilistic “white-box” approach allows us to analyze the impact of individual patterns on the classification and to compare the patterns that were found highly indicative to patterns found in related work.

In Hearst’s early work on pattern-based hyponym extraction (Hearst, 1992), very few hand-picked high-precision patterns were used to extract hyponyms from a text corpus. The recall was expectedly low. Many approaches to pattern-based extraction followed the idea of using a few hand-picked high-quality patterns. Most successful applications of these kinds of patterns are in extracting linguistic relationships.
One goal of this work was to broaden the pattern-base to achieve higher recall in extraction. Analysis of the patterns we found for different types of relationships confirmed the predictive power of many patterns that were identified by Hearst or that we would intuitively attribute to these relationships. For example, “⟨Subject⟩s, such as ⟨Object⟩” is a good indicator of a sub-class relationship. However, when these patterns are used for taxonomy induction, the ambiguity of this pattern becomes evident. Besides sub-class relationships, it can also indicate an occupation relationship, as in “modern artists, such as Picasso ...”. The algorithm also finds more domain-specific patterns. For example a pattern used in species classification: “⟨Object⟩ of the family ⟨Subject⟩”. Other patterns are domain independent and less predictive, but still appropriate, for example “⟨Object⟩s, including ⟨Subject⟩”.

Apart from sub-class/hyponymy-type relationships, we analyzed patterns for other types of relationships. The pattern “⟨Subject⟩ was born in ⟨Object⟩”, for example, always indicated the birthplace relationship. However, “⟨Subject⟩, the former president of ⟨Object⟩” and “⟨Subject⟩, prime minister of ⟨Object⟩” are also good indicators of a birth-place or nationality relationship. Not only do many countries require their presidents or prime ministers to be born in the country, most presidents and prime ministers were actually born in their country. The pertinence measure ensures that the likelihood that these patterns indicate birthplace are less affected by the occurrence of the same patterns in the president or prime minister relationships than a straightforward conditional probability computation would. This broadening of the semantics of a pattern beyond its immediate intension has usually been done using rule-based reasoning on top of the extraction, for example in SOFIE (Suchanek et al., 2009). The problem there is that rules need to be asserted manually and reasoned on separately, whereas in our case these implications are built into the extraction itself.

Other patterns that show a fairly high precision, are of the general form “⟨name⟩, ⟨location⟩ ⟨profession⟩”, as the examples of patterns and their significance to the birthplace
and occupation relationships in Table 5.6 show.

As described in (Kozareva et al., 2009), doubly anchored patterns, as in “⟨Subject⟩, american singer and ⟨Object⟩” perform with very high precision. The general form of these patterns for the occupation relationship is “⟨name⟩, ⟨location⟩ ⟨profession1⟩ and ⟨profession2⟩”. As the examples in Table 5.6 show, the location information can in this case be omitted without harming the indicative power of the pattern.

Table 5.6: Probability that a pattern indicates the birthPlace or occupation relationship

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Pattern</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthplace</td>
<td>⟨Subject⟩, ⟨Object⟩ comedian</td>
<td>0.7344144</td>
</tr>
<tr>
<td></td>
<td>⟨Subject⟩, ⟨Object⟩ actor</td>
<td>0.6413878</td>
</tr>
<tr>
<td></td>
<td>⟨Subject⟩, ⟨Object⟩ composer</td>
<td>0.5161507</td>
</tr>
<tr>
<td></td>
<td>⟨Subject⟩, ⟨Object⟩ guitarist</td>
<td>0.5161426</td>
</tr>
<tr>
<td>occupation</td>
<td>⟨Subject⟩, english ⟨Object⟩</td>
<td>0.7362891</td>
</tr>
<tr>
<td></td>
<td>⟨Subject⟩, russian ⟨Object⟩</td>
<td>0.7135882</td>
</tr>
<tr>
<td>occupation, doubly</td>
<td>⟨Subject⟩, american singer and ⟨Object⟩</td>
<td>1.0</td>
</tr>
<tr>
<td>anchored</td>
<td>⟨Subject⟩, * singer and ⟨Object⟩</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>⟨Subject⟩, * comedian and ⟨Object⟩</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Some patterns, on the other hand, are too weak to provide a classification into specific relationships by themselves. Such patterns indicate high-level property types that may not even be in the training corpus. For example, the pattern “⟨Subject⟩ from ⟨Object⟩” generally indicates a relationship of spatial or conceptual origin. Such a general pattern has little impact on the precise detection of a relationship type, but can indicate a general direction for the classifier.

Broadening the pattern base ensures better recall on the one hand and more confidence in the extracted relationships on the other. Hearst patterns are often used for taxonomy induction, but can miss the point, because hypernym/hyponym relationships do not always translate into superclass/subclass relationships. Often we want to instead extract an occupation relationship or a family or genus relationship in biology domains. The combination of many patterns allows a finer-grained classification and the pertinence measure allows multiple classifications with high certainty. This makes it possible to build classifiers for
large numbers of relationship types.

5.4.8 Discussion
In this IE work we explore how surface-pattern-based extraction can be improved by elevating the extraction to a concept level, rather than staying at the term-level. This approach poses the challenge of resolving ambiguous concept identifiers, but also gives the reward of being able to map extracted facts to formal domain models. Technically, the approach allows for efficient processing of textual data, especially when the sources are already indexed and patterns can be pulled from term-position vectors rather than from raw text. The major challenge of the approach from a machine-learning point of view is that we assume no negative training examples and at this stage also no user intervention to correct the algorithm in case of misclassified instances. This is addressed using statistical techniques that detect relationship boundaries and similarities. A drawback when using a pattern-based representation with a maximum-length pattern is that we miss out on information expressed in longer phrases. However, (Wu and Weld, 2010) shows that even in a parsing-based systems that uses a parse-tree-generalization algorithm, the extraction accuracy decreases with longer sentences.

5.5 Model Completion - Combining Definition and Description

This section describes the creation of a connected domain model as a combination of Domain Definition (Section 5.3) and Domain Description (Section 5.4). After a domain hierarchy is created based on a keyword description and possibly modified to best reflect the user’s expectations, the individual concepts in the hierarchy (i.e. the leaf nodes or instances) are listed and paired. Recall that this work regards fact extraction
as a classification task from concept pairs into relationship types. The pairing of concepts from the hierarchy can be done exhaustively, which will result in high processing cost when searching for patterns that express the concept pairs, because $n^2$ concept pairs will have to be considered given that $n$ concepts are in the model. Thus it is beneficial to restrict the concept pairs that are considered for fact extraction to those that are likely related.

5.5.1 Concept Pairing heuristic - Wikipedia

For domain hierarchies extracted from Wikipedia, a simple heuristic can be used to select a set $CP$ of concept pairs in which each pair could be related by one of the trained relationship types. We can assume that if two concepts are directly related, the corresponding Wikipedia pages will be linked. Equation 5.30 formalizes this idea. There, a concept pair $\langle C_i, C_j \rangle$ are concepts from the set of domain concepts $C$; $a_C$ represents a Wikipedia article that describes the concept $C$ and $\text{link}(a, b)$ means that a link exists from article $a$ to article $b$.

$$CP = \{C_i, C_j \in C| \exists a_{C_i}, a_{C_j} \in A, \text{link}(a_{C_i}, a_{C_j})\} \quad (5.30)$$

5.5.2 Concept Pairing heuristic - General Case

When the concept definition is extracted from a corpus that does not have unambiguous concept links, a pre-selection of the set $CP$ of concept pairs can be done using co-occurrence analysis on the text corpus that is used. Equation 5.31 shows a possible pre-selection strategy using pointwise mutual information (PMI). An outcome of $pmi(C_i, C_j) > 0$ indicates that $C_i$ and $C_j$ are dependent. The higher the threshold $\epsilon_{pmi}$ is set, the more dependent $C_i$ and $C_j$ will be and hence the more likely they will be related. The joint probability $p(C_i, C_j)$ that is part of the standard PMI definition can be computed by counting co-occurrence of terms that denote $C_i$ and $C_j$ in each document in the corpus.
\[ CP = \{C_i, C_j \in C | (C_i \neq C_j) \land pm(C_i, C_j) > \epsilon_{pmi}\} \] (5.31)

Different from the Wikipedia case, free text only contains concept mentions and does not reference the concepts directly. For this reason, PMI is computed over the co-occurrence of identified concept labels in a set of documents \( D = \{d_1, ..., d_n\} \). Equation 5.32 shows the PMI computation for this case. \( L_C \) means therein that the label \( L \) is a possible concept denotation for \( C \). Given that only patterns of limited length play a role in the IE algorithm, it is appropriate to only consider occurrences of terms that denote \( C_i \) and \( C_j \) if they are within a small window of tokens, that is, if the distance between the labels that denote \( C_i \) and \( C_j \) in document \( d \) is less than a threshold \( t \).

\[
pm(C_i, C_j) = \log \frac{|D| \cdot |\{d \in D: (L_{Ci} \in d) \land (L_{Cj} \in d) \land \text{dist}(L_i, L_j) < t\}|}{|\{d \in D: L_{Ci} \in d\}| \cdot |\{d \in D: L_{Cj} \in d\}|} \tag{5.32}
\]

### 5.5.3 Model-Creation

Section 5.4 described the general algorithm for extracting new facts. To connect a taxonomy or a domain hierarchy with named relationships, the text corpus is searched for occurrences of the concept pairs that were found using Equation 5.30 or 5.31 in conjunction with patterns that were learned during training. The \( CP2P \) matrix is built from these occurrences of concept pairs with patterns. Creating \( CP2P \) by applying Equation 5.26 and then performing the matrix multiplication from Equation 5.28 on \( CP2P \) and a trained \( R2P \) matrix yields the probabilities for relationships between the considered concept pairs.
5.6 Evaluation

This section presents evaluations for the automatic hierarchy creation part, the fact extraction part and their combination into completing a full domain model.

The algorithms are evaluated extensively by analyzing the results of the hierarchy creation and the Information Extraction parts separately and then evaluate the connected domain models:

1. Quantitative evaluation of the domain hierarchy extraction (Section 5.6.1)
   - wrt. other tools that extract terms or concepts relevant to a domain.
   - wrt. the MeSH hierarchy.

2. Quantitative evaluation of the facts extracted by the pattern-based method wrt. DB-Pedia and UMLS gold standards (Section 5.6.2)

3. Qualitative evaluation of the resulting connected domain models (Section 5.6.3)

5.6.1 Hierarchy Creation Evaluation

Guarino (Sure et al., 2004) suggests to compare a new ontology to a canonized domain conceptualization and then measure precision and recall by determining how well the ontology covers the conceptualization. Conceptually, this is a good method to determine ontology quality, but it causes practical problems, because when we use such a conceptualization as a gold standard to test the performance of the system, the problem of mapping between concept descriptions in both has to be resolved. Moreover, in many domains in which automatic extraction of ontologies or domain models would be desirable, no ready-made conceptualizations exist that could be used as a gold standard.

We decided to evaluate the extracted hierarchies qualitatively and quantitatively. In the qualitative analysis we manually compare the outcome of a focus query against the outcome
of the same (or similar, but optimized) query to a comparable service. In the quantitative analyses, we compare different generated hierarchies against gold standard taxonomies and glossaries. For these analyses we chose to create models from the financial domain and the biomedical domain, where we can compare to a trusted taxonomy in the form of the MeSH (Medical Subject Headings) (Rogers, 1960) hierarchy.

In all these evaluation methodologies we only evaluate the presence of concepts/entities in the hierarchy, not their placement in the hierarchy. The reason is that the extracted hierarchy is copied from the Wikipedia category hierarchy and thus only as good as Wikipedia itself.

5.6.1.1 Comparable services

To measure the quality of the results produced by Doozer, we created domain taxonomies and compared them with tools specialized in mining Wikipedia and human-composed glossaries. These comparisons are now described.

Sets by Google Labs (Google-Labs, 2004): The (now discontinued) service allowed the user to input between one and five example concepts. When comparing with domains created by our method using really short seed queries, such as “mortgage” or “database” or “federal reserve”, we found Google Sets worked better if that seed was repeated as an example five times rather than just once with four null inputs. Sets was also the only service we used in our comparisons that could not be restricted to return results that lie within the Wikipedia name space.

Grokker by Groxis, Inc.: This (now discontinued) service allowed the user to find and organize related concepts, and can be constrained to return only Wikipedia concepts. We invoked the service using our seed query, with the exception of introducing white space and explicit OR keywords where necessary.

PowerSet is a service to mine Wikipedia using either simple queries or natural language questions. Where Power Labs returned better results using a question rather than a seed
query, as was the case for instance with “What about mortgage” versus “mortgage”, we used the input that produced higher-quality results.

Finally, we also compared against results obtained using **Wikimedia search**, i.e. the default Wikipedia search option, which takes our unmodified seed queries. Since a full-text search is the first step in creating domain hierarchies, this comparison gives a good idea of the information gain we get using the graph expansion step.

### 5.6.1.2 Qualitative comparison of top-10 ranked results

Given a domain taxonomy, our analysis proceeds by comparing the different lists of terms generated for the same query by the above mentioned services, first against each other and then against a standard reference. Table 5.7 shows the top ten results returned against the query “mortgage”. Some services show rapid deterioration even in the first ten results. For instance, while finding gems such as “houses for sale” and “personal finance”, Google Sets also finds irrelevant concepts such as “photos” and “overview”. Likewise, Grokker brings up several loosely related terms such as “Tort law” and the very generic but relevant term “Bank” among its top ten hits. Powerset does well to find “2007 Subprime mortgage financial crisis,” perhaps from a direct keyword hit, and Wikipedia Search likewise brings up several keyword hits. Doozer also brings up some loose hits such as “UK Mortgage Terminology”, but does rather well to find “interest-only loan”. Doozer and Grokker do well to find more than just keyword hits, while still keeping good precision.

### 5.6.1.3 Quantitative comparison of tools against a reference taxonomy

The ideal reference list should be agreed upon by domain experts and disambiguated with other domains. In what follows, we interpret terms with high relevancy as those which are put into the domain by agreement among domain experts and have negligible, if any, ambiguity with other domains. In absence of an ideal reference list, the reference lists used
Table 5.7: Comparing the top ten ranked results returned by our method against four competing methods

<table>
<thead>
<tr>
<th></th>
<th>Google Sets</th>
<th>Grokker</th>
<th>Doozer</th>
<th>Powerset</th>
<th>Wikimedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>mortgage</td>
<td>Tort law</td>
<td>Mortgage Loan</td>
<td>Mortgage loan</td>
<td>Mortgage</td>
</tr>
<tr>
<td>2nd</td>
<td>overview</td>
<td>Lost, mislaid, and abandoned property</td>
<td>Mortgage</td>
<td>Mortgage broker</td>
<td>Mortgage loan</td>
</tr>
<tr>
<td>3rd</td>
<td>contact brochures</td>
<td>Mortgage loan</td>
<td>Teachers' Building Society</td>
<td>Mortgage underwriting</td>
<td>Mortgage-backed security</td>
</tr>
<tr>
<td>4th</td>
<td>photos</td>
<td>Law of Property Act 1925 (c.20...)</td>
<td>Mortgage-backed Security</td>
<td>Subprime mortgage financial crisis</td>
<td>Mortgage discrimination</td>
</tr>
<tr>
<td>5th</td>
<td>personal finance</td>
<td>Leasehold estate</td>
<td>Agency Securities</td>
<td>Mortgage discrimination</td>
<td>Mortgage discrimination</td>
</tr>
<tr>
<td>6th</td>
<td>foreclosures</td>
<td>Non-possessory interest in land</td>
<td>Uk Mortgage Terminology</td>
<td>Shared appreciation mortgage</td>
<td>Mortgage broker</td>
</tr>
<tr>
<td>7th</td>
<td>houses for sale</td>
<td>Bank</td>
<td>Mortgage Note</td>
<td>Mortgage insurance</td>
<td>Mortgage bank</td>
</tr>
<tr>
<td>8th</td>
<td>real estate training</td>
<td>Non-recourse debt</td>
<td>Adjustable Rate Mortgage</td>
<td>Biweekly Mortgage</td>
<td>Adjustable rate mortgage</td>
</tr>
<tr>
<td>9th</td>
<td>people search</td>
<td>Lateral and subjacent support</td>
<td>Graduated Payment Mortgage Loan</td>
<td>Mortgage</td>
<td>Lenders mortgage insurance</td>
</tr>
<tr>
<td>10th</td>
<td>mortgage lending tree</td>
<td>Mortgage</td>
<td>Interest-only Loan</td>
<td>Foreign currency mortgage</td>
<td>Ameriquest Mortgage</td>
</tr>
</tbody>
</table>

in this work are formed from domain glossaries of terms. To further measure the accuracy with which the tools discover relevant terms from the same training corpus, i.e., Wikipedia, we have also correlated the terms in the domain glossaries with the list of terms produced from a topic search of the Wikipedia corpus. In our analysis, we used a glossary (Wheeler and Wheeler, 2010) of financial terms which has been pre-categorized into domains. In particular, we utilized the list of terms in the federal reserve and mortgage domains. The tools from Google, Grokker, Powerset, and Wikipedia as well as ours, were each queried with each of these two seeds to produce two domain lists per tool. Since the financial glossary regards all its terms as equally relevant, so did we drop any weights, probabilities and ranking computed by the tools; thus, all of the terms, appearing in each tool’s list, are equally relevant within that list. In order to reduce the terms from the glossary down to only the ones found in a search for the respective seed topics in Wikipedia, we produced the reference list as the intersection of the respective glossary and Wikipedia search results.
These reference lists then contain terms that the author of the glossary would consider relevant to the respective domains and that are also present in the corpus upon which the taxonomies are built. The values of the $F_1$ measure (Lau, 2007) were then computed for the lists generated by the tools. The results are illustrated in Figure 5.7. We note that Doozer’s results in terms of the $F_1$ measure are at least a factor or two improvement over those of the other tools. This difference in performance can be attributed to the amount of noise in the topic search results of Wikipedia. In order to quantify this further, we considered the fact that there might be some relevant terms which appear in the topic search results from Wikipedia but do not appear in the standard glossary. Thus, we compute the fraction of relevant terms found in a topic search of Wikipedia as

$$v = f + (1 - f) \frac{|W \cap G|}{|W|}$$  (5.33)

where $W$ refers to the set of search results from Wikipedia, $G$ is the set of terms in the glossary, and $f$ is the fraction of terms from any snapshot of the complement of $W \cap G$ in $W$ that are deemed to be relevant to the domain. Using the “mortgage” domain as a test case, we formed the disjunction of the lists from the glossary and search results from Wikipedia. We then took a random snapshot of the terms in the disjunction and then found those terms that were present in the search results from Wikipedia and relevant to the mortgage domain; this gave us an estimated value for $f$. The stemmed terms from the snapshot are listed in Table 5.8.

Using Equation 5.33, the percentage of the search results in Wikipedia that are relevant is estimated to be 32%. This suggests that those tools that rely primarily on the topic searches of Wikipedia will include a high percentage of noise. As further evidence of this, the precision and recall of the terms in the glossary against the search results of Wikipedia as the reference list are 0.14 and 0.04, respectively. Considering that precision and recall are measured against the search results of Wikipedia, the very low recall by the glossary
is indicative of a large percentage of irrelevant terms in the search results of Wikipedia. Likewise, the low precision for the glossary is indicative of a large percentage of relevant glossary terms that are not present in the search results of Wikipedia. These results provide evidence that the use of topic search results of Wikipedia will have high rates of both false positives and false negatives if they were used as the sole basis for taxonomies. As described above, the approach reported in this work increases the recall primarily by exploiting the link structure of Wikipedia to find additional topics that are similar to an initial set of topics. Furthermore, we use domain relevancy statistics (weights and conditional probabilities) to prune intermediate lists, thereby increasing the precision of Doozer’s results, as evidenced by the results in Figure 5.7.
Figure 5.7: $F_1$ measures, computed against a reduced glossary, for the lists of terms generated by various mining tools.

Figure 5.8: The Oncology subtree of the created MeSH-related category hierarchy. The size of the rectangles indicates the number of descendants. Lighter rectangles indicate categories, darker rectangles indicate individuals.

### 5.6.1.4 Comparison against MeSH

Comparing the generated list of domain terms to a Gold Standard such as MeSH (Medical Subject Headings) allows us to get a quantitative evaluation of the extraction quality in a scientific field. However, this kind of comparison is biased towards the concepts in the gold standard. Terms in the extracted model that are relevant, but not in the gold standard will not count towards the model’s precision and recall. Domain ontologies and glossaries usually contain terms for immediate domain concepts rather than terms that are highly indicative of a domain. The term “cancer”, for example, is very important for, but not highly...
indicative of the oncology field. The content of the created domain models are meant to be used in retrieval and classification tasks. For this reason, finding the phrase “MIT Center for Cancer Research” in an article is very indicative of the oncology domain, but not useful in a specific biomedical taxonomy such as MeSH. Nevertheless, in order to have a numerical evaluation of the ontology creation process, an automated Gold-Standard evaluation (Brank et al., 2005) is performed. We extracted all MeSH terms in the Neoplasms subtree to compare them against an automatically generated Neoplasms domain model. For the extraction, the algorithm was run with the following description:

- **Focus description**: (Adenoma Carcinoma Vipoma Fibroma Glucagonoma Glioblastoma Leukemia Lymphoma Melanoma Myoma Neoplasm Papilloma) AND (medicine medical oncology cell disease)

- **Domain categories**: oncology, medicine, types of cancer

- **World view**: Biology

An abstract view of this domain model is shown in Figure 5.8. The visualization is done using the Jambalaya environment. Using the size of the boxes as indicators of the relative number of descendants, it is apparent that the “types of cancer” category gets the largest weight in the domain model. In fact, out of a total of 222 extracted instances, 135 belong to the category “types of cancer”.

Just like Wikipedia, MeSH is constantly evolving, albeit the changes are performed by domain experts only. In order to show how useful an automatic extraction of a domain model can be to stay up-to-date without investing human effort, the extracted domain model is compared against the Neoplasm subtrees of the MeSH versions from the years 2004 and 2008. Alignment of Wikipedia and MeSH is not in the scope of this work, therefore we only evaluate string matches between the extracted domain model and the Neoplasms subtree. The terms in the generated domain model are matched against two subsets of both MeSH Neoplasms versions. (1) is the full set of terms, (2) is the subset of MeSH terms that can
Table 5.9: Terms in the Neoplasm subtree of MeSH. Column (1) contains the number for all terms, column (2) contains the number of terms that could also be found on Wikipedia.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>405</td>
<td>147</td>
<td>36.3</td>
</tr>
<tr>
<td>2008</td>
<td>636</td>
<td>227</td>
<td>35.7</td>
</tr>
</tbody>
</table>

Table 5.10: Parameter Settings for three different experiments to reproduce the MeSH Neoplasms subtree.

|       | Search Results | Expansion Threshold | min p(Domain|Article) |
|-------|----------------|--------------------|-------------|
| 1     | 40             | 0.5                | 0.1         |
| 2     | 40             | 0.5                | 0.4         |
| 3     | 25             | 0.8                | 0.5         |

Figure 5.9: Evaluation wrt. the Wikipedia subset of terms in MeSH versions of 2004 and 2008.

...actually be found in the Wikipedia titles and their synonyms and is thus the maximum number of matches we can possibly achieve with the current method. See table 5.9 for an analysis of the overlap between MeSH-Neoplasms and Wikipedia.

We performed the comparison with the MeSH C04 Neoplasms subtree using three different experimental setups for domain hierarchy generation; ranging from more precision-oriented to more recall-oriented. By changing the thresholds in various steps in the algorithm as seen in Table 5.10, we achieve more expansion or more reduction. Recall oriented
means using a lower search threshold (more initial results), a lower expansion threshold (more similar nodes) and a low domain-importance threshold (fewer nodes deleted because of conditional probability). Precision-oriented means that higher thresholds were set. We evaluated the created domain hierarchies with respect to the MeSH versions of 2004 and 2008. We achieve a precision of up to 48% and a recall of up to 72% wrt. MeSH 2008 as well as precision of up to 34% and recall up to 76% wrt. MeSH 2004. Figure 5.9 shows the results of this evaluation. One reason for not achieving higher scores can be seen in the different scopes of Wikipedia and MeSH, another lies in the different goals that we have for the hierarchy creation application compared to the intention of MeSH. We want terms that are highly indicative of a domain, not necessarily those that domain experts use within their domain, but that are ambiguous outside the domain. MeSH, on the other hand, is clearly meant for the categorization of biomedical concepts regardless of their indicative power for a classifier. The $\nu$-Measure introduced in Equation 5.33 accounts for this difference in scope by incorporating an estimated factor of correctly extracted concepts that were not found in the gold standard. The estimates from the $\nu$-Measure results are also included in Figure 5.9 and range between 64% and 74% wrt. MeSH 2004 and between 71% and 75% wrt. MeSH 2008. The $\nu$-Measure can give us a good idea of the hierarchy’s utility, because it considers both matched concepts as well as unmatched but nevertheless correct and relevant concepts.

Even though the scope of the extracted hierarchy is different from the gold standard, the focus was the same. Out of a total of 222 extracted instances in the high-precision model (Experiment 3), 135 belong to the category types of cancer. Other categories and terms that are relevant to the neoplasms domain can also be found, such as radiobiology and therapy, which, amongst others, share the instance Radiation therapy as well as Chemotherapeutic agents, Tumor suppressor gene and Carcinogens. These additional results that are not part of MeSH exemplify the difference in scope between Doozer’s domain hierarchy and MeSH. Whereas the MeSH C04 subtree restricts itself to listing different
types of neoplasms, Doozer discovers many related concepts that are important for a general classification model of the neoplasms domain, but do not represent types of neoplasms themselves. We take this as a strong indication that Doozer performs well in the task of producing domain hierarchies even in such a specialized domain as the neoplasms field.

5.6.2 Fact Extraction Evaluation

This section presents an evaluation of the pattern-based IE algorithm as a whole as well as the pertinence computation in particular. The evaluation is done with respect to two gold standards that are described in section 5.6.2.1. Any evaluation with respect to gold standards with only positive examples will have the disadvantage of only showing true and false positives. However, it provides an analysis with a statistically significant number of examples.

5.6.2.1 Datasets

The IE approach taken in this work applies to any combination of corpora, given that there is a training corpus with facts (triples or other representation) that instantiate different relationship types and a text corpus that has textual manifestations of many of the facts, so textual patterns can be learned. To demonstrate the applicability of the algorithm in different domains, the following two combinations of fact and text corpora were chosen:

- DBPedia Infobox fact corpus and Wikipedia text corpus
- UMLS fact corpus and Medline-abstracts text corpus

The contents of both dataset combinations are very different. Wikipedia is an encyclopedia that provides a broad cross-section of human knowledge with currently about 4 Million articles. The DBPedia Infobox corpus (Bizer et al., 2009b) contains mostly political, geographical, biographical and entertainment-related facts. Medline is a collection
of over 18 million biomedical publications. The abstracts to these publications are freely available and are used in this work. UMLS (Lindberg et al., 1993) is a collection of facts from more than 100 different controlled vocabularies and databases in the biomedical field. It covers taxonomic and meronomic relationships as well as relationships describing drug interactions, disease-sites and many more. The scope of UMLS is much narrower, but a lot deeper and more detailed than that of the DBPedia Infobox corpus. The same holds for the text corpora. Wikipedia contains a broad collection of general knowledge articles whereas Medline contains focused and detailed scientific publications. Neither Wikipedia nor the MedLine abstracts were significantly preprocessed with the exception that all Wiki-specific syntax, links, Infoboxes and category assignments were stripped from the text to assure that all pattern representations were taken from raw text rather than from a structured representation.

In both cases, the fact corpus and the text corpus are only loosely connected, insofar as it is known that many of the fact triples are expressed in the form of raw text. Using loosely connected corpora for supervised training has recently been termed Distant Supervision (Mintz et al., 2009). This loose coupling between corpora brings about advantages and disadvantages. On the one hand it allows us to access unprecedented amounts of training data in the form of facts that were asserted either by a community of experts in the case of UMLS or a broad community of experts and laypersons in the case of DBPedia. These datasets are likely to keep growing and thus provide more and more training facts. On the other hand, since the text is not annotated, it is uncertain that the features that are garnered from the text corpus are representations of the facts in the fact corpus.

Since the objective of fact extraction in this dissertation is extracting named relationships between entities/concepts in a hierarchy, datatype properties were removed. Moreover, only types of relationships that were instantiated by a minimum of 25 fact triples were taken into consideration. In our opinion, this does not violate the premise that the datasets used should not be pre-processed, because the selection is based on fixed criteria, not on
the quality of the facts or their availability in the corresponding text corpus.

5.6.2.2 Gold Standard Evaluation

For this evaluation all patterns for all distinct Subject-Object pairs were extracted in both corpora (DBpedia Infobox and UMLS). The resulting matrices were randomly split into 80% training examples and 20% testing examples. Training and testing sets were completely disjoint, i.e. no subject or object that appeared in one set was allowed to be in the other. The splitting was repeated 10 times and the results averaged. The size of a testing set was about 11,000 facts in the DBPedia case and about 4100 in the UMLS case. Precision and recall were first computed on a per-relationship basis and then averaged over all relationships. This normalized evaluation yields lower values, because relationships with many examples tend to perform with higher precision and recall. Due to the normalization, these types of relationships are counted equally to relationship types with very few examples, thus diminishing the impact of each correctly extracted fact from an abundantly represented relationship. However, for the evaluation of an open IE approach this is more meaningful than averaging over all extracted facts, because it shows that even higher numbers of relationship types do not break the system.

Figures 5.10 and 5.11 show precision and recall with respect to the gold-standards DBpedia and UMLS. The horizontal axis indicates the confidence threshold $\epsilon_{rel}$ that was used, i.e. only statements that were extracted with a probability greater than $\epsilon_{rel}$ were taken into consideration. The average values show the arithmetic mean precision and recall values over all relationship types, the max values show the maximum precision and recall among the relationship types. Precision and recall are thereby computed according to Equations 5.34 and 5.35.
Figure 5.10: Precision and recall: DBPedia test set and Wikipedia text corpus.

\[
\text{Precision}_{\epsilon_{rel}} = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \frac{|\text{correct facts for } R, \text{ confidence } > \epsilon_{rel}|}{|\text{all extracted facts for } R, \text{ confidence } > \epsilon_{rel}|} \tag{5.34}
\]

\[
\text{Recall}_{\epsilon_{rel}} = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \frac{|\text{correct facts for } R, \text{ confidence } > \epsilon_{rel}|}{|\text{all gold standard facts for } R|} \tag{5.35}
\]

Figure 5.10 shows the automatic evaluation of precision and recall over all cross-evaluation sets of the DBPedia-Wikipedia corpus. Considered were only those relationship types for which more than 25 possible occurrences were found in the Wikipedia corpus, which amounted to an average of 107 distinct types. Only direct hits in first rank according
Figure 5.11: Precision and recall: UMLS test set and MedLine text corpus.

to Equation 5.36 were taken into account.

\[ R_{CP_i} = \arg \max_j p(R_j|CP_i) \text{ if } p(R_j|CP_i) > t \] (5.36)

The evaluation shown in Figure 5.11 over the UMLS-MedLine corpus is analogous. The curves are in general steeper than in the DBpedia-Wikipedia case and go up to over 80% precision as the confidence threshold increases. A pattern analysis showed that in a scientific corpus the expressions are more specific. This translates to more specific patterns that apply to fewer concept pairs. The precision and recall lines cross at comparable points in both cases, which indicates that there is a baseline of patterns that describe more general types of relationships. In all cases the results were well over the random baseline. Even in the high recall regions the average precision is at least 35% and goes up to 65% as the confidence threshold increases, with some relationship types showing perfect precision. The high recall especially in the DBPedia case is a definite improvement over the
approaches mentioned in the related work section that are more precision-oriented. It can thus be shown that with a basic probabilistic approach the surface pattern analysis can be used to connect and augment domain models in information retrieval applications.

To show the difference in performance when changing the evaluation style, in the following evaluation the precision was averaged over all extracted facts without first averaging over the relationship types. Figure 5.12 shows these results. It is apparent that the precision increased significantly, especially in the DBPedia case. The reason is that on average relationship types with many examples tend to perform better than those with fewer examples. Figure 5.13 shows this trend. This figure was generated by ordering the <Number of example triples, Relationship-type precision>-tuples by precision and plotting a trend-line averaged over a window of 20 tuples. Using an oversampling strategy for underrepresented types of relationships rather than random sampling would remedy this discrepancy for evalu-
Figure 5.13: Support vs. precision of relationships. The left scale indicates the training examples, the right one indicates precision for a relationship type.

uation purposes. However, when looking at LOD as constantly streaming data, rather than a static corpus, it is more realistic not to make assumptions about the distribution of relationship types, because the distribution changes depending on editing dynamics in the community. In an application scenario the classifier would naturally be trained on all available facts to have a maximum gain of pattern probability information at each point in time and thus it is bound to the distribution of relationships types in the dataset that it is dealt.

5.6.2.3 Comparison to other approaches

As outlined in the related work section, the information extraction approach taken here makes different assumptions concerning the training data and uses different strategies in processing the data and in training classifiers than its competitors. It is thus difficult to make a fair comparison. However, WOE (Wu and Weld, 2010) and the work by Mintz et al. (Mintz et al., 2009) are the most likely comparisons insofar as WOE trains on Wikipedia
Infobox relationships and one of our experiments is also restricted to this dataset. Mintz et al. performed experiments using POS-tagged patterns, as well. To get a fair comparison with WOE, the three high-precision types of relationships *spouse*, *occupation* and *birthplace* are used in this evaluation. $WOE_{POS}$ learns classifiers from POS-tagged sentences, $WOE_{Parse}$ from fully parsed sentences. Figure 5.14 compares the average of these results with both WOE flavors. $WOE_{POS}$ achieves high precision in the low-recall regions, Doozer++ maintains a better precision as the recall increases and it achieves overall higher recall than $WOE_{POS}$. $WOE_{Parse}$ performs better, maintaining high precision across all recall values. This shows that there is still a gap between surface-pattern-based and NLP-based approaches, but Doozer++ succeeded at narrowing this gap. The comparison to Mintz (see Figure 5.15) is performed by averaging across all 107 relationships (Mintz uses 102). For this comparison, a more precision-oriented and a more recall-oriented experiment were performed. In the precision-oriented run the patterns were not generalized, whereas in the recall-oriented run a generalization was performed with up to two wild cards per pattern. The precision-oriented experiment peaks at a higher precision than Mintz, but has a sharper drop until it maintains a higher precision in the high recall regions as well. The recall-oriented run has a much smoother curve, but starting at a lower precision than Mintz. It however maintains a higher precision starting at a recall of about 0.2. Mintz et al. use a “multi-class logistic classifier optimized using L-BFGS with Gaussian regularization”. The results suggest that, even with a weaker pattern representation, our classifier using conditional probabilities modified through entropy and pertinence computations, outperforms this well-regarded ML algorithm.

Both comparisons against WOE and Mintz show that Doozer++ generally outperforms
other non-NLP techniques, even when they use POS-tagging and it achieves a narrowing of the gap between NLP and non-NLP techniques.

The extractions done with WOE are available for download. It appears that a triple is counted as correct, when the sentence the triple occurs in is correctly split. However, many of the extracted statements are not informative, because e.g. the subject is a pronoun, some other coreference, such as “those workers”, compounds, such as “Shares of the New York-based company” or phrases that only reference in the proper context, such as “six of 11 countries” or “any of the myriad collaborators”. This results in triples such as “any of the myriad collaborators → may leave → a brilliant mark”, which only has meaning in a proper context. Predicates are normalized mostly to the extent that verbs are stemmed, but there is no alignment of synonymous predicates to assure that the predicate represents a formal relationship. Since Doozer++ is concept-centric, these kinds of non-referential subjects and objects are not extracted. All triples have proper concepts or entities as subject and object, not just concept designators. Due to the supervised classification, the predicates of the triples reflect formal relationships. Hence triples that are counted as correct are also formally valid and computationally usable.
Figure 5.14: Precision/Recall curve for select relationships, comparison between Doozer++ and the POS-version as well as the parse-version of WOE.

Table 5.11: Criteria, queries and results for the Domain Definition part of Model Creation

<table>
<thead>
<tr>
<th>Domain</th>
<th>Criteria</th>
<th>Query</th>
<th># concepts</th>
<th>Precision Domain Def</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Web</td>
<td>Global, Compact, Deterministic, Lasting</td>
<td>“Semantic Web” OWL, Ontologies RDF</td>
<td>143</td>
<td>0.98</td>
</tr>
<tr>
<td>Harry Potter</td>
<td>Global, Compact, Deterministic, Lasting</td>
<td>“Harry Potter” dumbledore griffindor slytherin</td>
<td>134</td>
<td>0.98</td>
</tr>
<tr>
<td>Beatles</td>
<td>Global, Compact, Deterministic, Lasting</td>
<td>Beatles “John Lennon” gryffindor slytherin</td>
<td>250</td>
<td>0.99</td>
</tr>
<tr>
<td>India-Pakistan Relations</td>
<td>Local, Loose, Unexpected, Lasting</td>
<td>India Pakistan Kashmir</td>
<td>129</td>
<td>0.99</td>
</tr>
<tr>
<td>US Financial crisis</td>
<td>Local, Compact, Unexpected, Transient</td>
<td>tarp “financial crisis”, “toxic assets”</td>
<td>146</td>
<td>0.93</td>
</tr>
<tr>
<td>German Chancellors</td>
<td>Local, Compact, Deterministic, Lasting</td>
<td>“German chancellor”, “Angela Merkel” “Helmut Kohl”</td>
<td>124</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 5.12: Seed-terms excerpt for the Human Cognitive Performance model

<table>
<thead>
<tr>
<th>Cognitive Psychology</th>
<th>Neuroscience-conceptual</th>
<th>Neuroscience-functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstract reasoning</td>
<td>activated cortical volume</td>
<td>acute and chronic hypoxia</td>
</tr>
<tr>
<td>abstraction</td>
<td>activation of frontal cortex</td>
<td>adult neurogenesis</td>
</tr>
<tr>
<td>acquired skills</td>
<td>adaptability between internal configurations</td>
<td>biomarkers</td>
</tr>
<tr>
<td>acquisition of cognitive skills</td>
<td>adaptive neuronal plasticity</td>
<td>brain oxygenation level</td>
</tr>
<tr>
<td>affection</td>
<td>amygdala</td>
<td>brain glucose metabolic rate</td>
</tr>
<tr>
<td>analogy making</td>
<td>anterior cingulate cortex</td>
<td>brain metabolism</td>
</tr>
</tbody>
</table>
5.6.3 Evaluation of the full Domain Models

The full model creation is evaluated in two different application scenarios. The first one is geared towards rapid model creation for general purpose information filtering and browsing tasks, the second one is aimed at guided scientific information retrieval.

5.6.3.1 Models for Information Filtering and Browsing

To properly evaluate the full model creation for information filtering, we generated six models of different domains. To get a broad spectrum of models, we build them according to the semantic scope dimensions that were identified in (Purohit et al., 2011):

1. Global vs. Local

2. Compact vs. Loose

---

*A collection of triples extracted by TextRunner, WOE and Reverb and their evaluations can be found here: http://www.cs.washington.edu/homes/afader/data/reverb_emnlp2011_data.tar.gz*
Figure 5.16: Precision of the fact extraction for the on-demand created Models, depending on the confidence threshold $\epsilon_{rel}$.

Figure 5.17: Relative recall of the fact extraction for the on-demand created Models, depending on the confidence threshold $\epsilon_{rel}$.

3. Deterministic vs. Unexpected

4. Transient vs. Lasting

The models were created for the criteria shown in Table 5.11. The Query column shows the queries that were sent to the Doozer hierarchy creation. The models were then manually evaluated. The number of concepts that were contained in each model gives an idea of coverage. The consistently high precision for the domain definition shows that the concepts
The hierarchies were then given to the relationship extractor to connect the concepts. The facts in the resulting models were again manually evaluated. This evaluation was done leniently in the sense that facts were counted as correct if the relationship type was very close to a correct type. For example, the triple \( \langle \text{Helmut Kohl} \rightarrow \text{commander} \rightarrow \text{German Chancellery} \rangle \) was counted as positive, even though the chancellor is not a military commander of the chancellery. The recall values are taken relative to the number of correct facts extracted using the lowest confidence threshold. The actual recall as it pertains to the domain is difficult to estimate, because the number of possible facts in a domain is unbounded. Figures 5.16 and 5.17 show precision and relative recall of the fact extraction part for these models.

Figure 5.18 shows an excerpt of the model created for the relationships between India and Pakistan. As in all the other example models, none of the asserted facts was directly taken from DBpedia. By placing the term “Kashmir” into the query, Doozer++ put emphasis on that part of the India-Pakistan relations. The algorithm correctly identified Pakistani leaders. However, it does not distinguish between current leaders and former leaders. This
is due to the fact that training set from the DBpedia corpus also does not make this distinction. The extractor found many triples that give important contextual information, such as the fact that the second president of Pakistan, Ayub Khan, was born in (the former British) India or that Pakistan was founded by Muhammad Ali Jinnah. It also correctly identified the region of Kashmir as belonging to both countries, but placed the state of Azad Kashmir into Pakistan and Jammu-and-Kashmir into India.

5.6.3.2 Evaluation of the Cognitive Science Model

As mentioned in the introduction, we developed a larger domain model for the research area of Human Cognitive Performance. This model currently serves as background knowledge for a domain-centric semantic browser named Scooner (Cameron et al., 2010). Here, the objective was to have a comprehensive account of the domain with high precision and recall. For this reason, a domain expert provided us with 3 lists of terms that are important in the domain. These lists were from the domains of cognitive psychology and neuroscience, whereby the neuroscience terms were again split into conceptual and functional aspects. An excerpt of these lists can be found in Table 5.12. The seed query for the model creation was a conjunction of all lists, whereby the terms in each list were disjunct. The query thus looked as follows: (abstract reasoning OR abstraction OR ...) AND (“activated cortical volume” OR “activation of frontal cortex” OR ...) AND (“acute and chronic hypoxia” OR “adult neurogenesis” OR ...).

Since the goal of the corresponding project was to have guided browsing of MEDLINE articles using entities and relationships relevant to the domain of human cognitive performance, the embellishment of the hierarchy with facts that instantiate domain-relevant
relationships was critical. Here, precision was more important than in the previous examples, even though, as a browsing guideline, recall was still the primary focus. To connect the entities in the hierarchy with named relationships from the biomedical domain, the relationship model in the form of the $\mathbf{R2P}$ matrix was trained by extracting patterns from the MEDLINE abstracts text corpus based on facts available in the UMLS corpus. Figure 5.1 in the introduction showed a small excerpt of the connected model. Figure 5.20 shows a more strongly connected part of the extracted concept graph. Of particular interest for the cognitive science domain are the $\text{receives\_input\_from}$ and $\text{sends\_output\_to}$ relationships that indicate interaction of brain regions in this model. These relationships do not exist at all in DBPedia and whereas some of them can be found in UMLS, the UMLS concepts cannot always be mapped to the coarser Wikipedia concepts. Figure 5.19 shows the first five levels of the resulting category hierarchy in a Treemap representation. Spatial inclusion in the Treemap represents a subcategory relationship, the size of the rectangles indicates the number of descendants and hence the importance of the category to the domain.
We examined the extracted facts and removed any that were already available in the UMLS training corpus from the sample that was evaluated. Cognitive scientists and biologists at the Air Force Research Lab evaluated 415 extracted facts that had a confidence score of 0.7 or higher. The scientists scored each fact on a scale of 1 to 9; 1 being plain wrong and 9 being true, novel and interesting. Figure 5.21 shows the scoring. It displays the percentage for each score and cumulative percentages for scores 1-2 (incorrect: 21%) and 3-9 (correct: 79%) respectively. About 30% of the extracted facts was deemed novel and interesting. The scoring rationale is as follows:

7-9: Correct Information not commonly known
5-6: General Information that is correct
3-4: Information that is correct, but trivial
1-2: Information that is overall incorrect
5.6.4 Discussion

Analysis of the fact extraction shows that a high percentage of the facts deemed as novel and interesting were extracted based on highly specialized low frequency patterns. These patterns appear in the text corpus few times with the provided training facts and tend to appear with subject and object concepts that fall in the same domain classes or range classes respectively. Figure 5.22 shows the average score for extracted facts for each confidence score from 1 to 9. The figure shows that high-quality facts get highest scores, but also some incorrect facts were extracted with high confidence. This is because within the pattern distribution, those patterns that tend to identify particular relationships with high confidence are in the long tail of the pattern frequency distribution. However, noisy patterns also tend to occupy this space. With the absence of negative examples, these are not easily distinguished. An analysis of the facts that were incorrectly identified with high confidence shows that they largely fall in two categories. The first is that of a formally incorrect but metaphorically correct relationship or of generally very high relatedness. For example,
the extracted assertion \((\text{Interpeduncular Cistern} \rightarrow \text{disease has associated anatomic site} \rightarrow \text{Cerebral peduncle})\) is incorrect, because the Interpeduncular Cistern is not a disease. However, it does have the associated anatomic site \text{Cerebral peduncle}. The second is that of incorrect directionality. In many cases the asserted relationship is correct, but points in the wrong direction. For example, the assertion \((\text{Pituitary Gland} \rightarrow \text{sends output to} \rightarrow \text{Supraoptic nucleus})\) is incorrect because the supraoptic nucleus sends output to the pituitary gland, not vice versa. Often, when these directional relationships are described in text, the direction is expressed in the context rather than in the phrase that relates the two entities. A possible remedy for these kinds of errors that are caused by patterns incorrectly identified as high-quality even though they are either noise or belong to a different type of relationship is to retrain the classifier with these incorrectly classified statements as negative examples. The NELL project (Carlson et al., 2010b) uses this kind of active learning to improve its precision. For future work we will incorporate an active learning component, as well.

An interesting aspect of the analysis of incorrectly identified relationships in sec-

Figure 5.22: Distribution of extraction confidence across the expert scores. High-quality facts get highest scores, but also some incorrect facts were extracted with high confidence.
tions 5.6.3.1 and 5.6.3.2 is that many of them represent the correct relationship qualia (Cimiano and Wenderoth, 2005). A business merger is often referred to similarly to a marriage, holding political office is often described similarly to holding a military rank or being a military commander. Thus it makes sense to connect the merging companies with the spouse relationship or the chancellor to the chancellery with the commander relationship. Lakoff and Johnson (1980), amongst others, identified evidence that the use of basic metaphors in natural language is ubiquitous. Turney’s (2006) work on word pair analogies also provides statistical significance to these claims. When encountering analogous relationships such as the ones mentioned, more well-formalized background knowledge will allow us to rule out wrong types of relationships, when we know that e.g. the domain and range of the spouse relationship must be Person.
Knowledge Verification and Propagation

6.1 Introduction

The previous chapters laid the foundation for a model-centric, i.e. user-interest-centric extraction of knowledge. In this chapter it is shown how to use the application of the extracted facts and their implicit and explicit evaluation in a hermeneutic circle of knowledge acquisition. This means, use and evaluation are an intrinsic part of knowledge generation.

It can be assumed that any kind of user behavior on a web site tells something about the information content of the site with respect to the user’s goal. To be able to collect user information during searching and browsing, Scooner (Cameron et al., 2010; Kavuluru et al., 2012) was developed, a browser that allows trail-blazing, i.e. following information trails by browsing of facts that are dynamically added by the browser instead of statically by the content-creators. Scooner allows the collection of search and click-stream data and it lets users vote directly on the facts (triples) that are the building blocks of the information
trails.

In this example scenario the automatically created domain models are used to search and browse content. Figure 6.1 shows an example search for the concept *dopamine receptor* restricted to a MedLine dataset and using a model of human cognition to focus the search. Terms that have been found as subjects or objects of triples are highlighted. When the user hovers over the highlighted terms, a context menu with a list of relationships opens up that shows how the concept described by the term can be related to other concepts. After choosing a relationship type and an object concept (or subject concept for reverse exploration of triples), a list of target web pages is shown that are relevant to the Subject-Relationship-Object triple chosen by the user. In the example the user followed a trail of research articles involving 2 triples. Note that the triples themselves may not be connected, but the connection is given in the research paper that was the target of the first triple and the source of the second. In these cases the user implicitly creates a relation between these two previously unrelated triples. An evaluation of this implicit creation of a semantic link is left to future work, because we are as of yet lacking sufficient data.
6.1.1 Explicit Validation

The top right corner of Figure 6.1 shows the trails that have been followed to find the literature that was sought. The user can approve or disapprove of each statement. This voting is logged and the facts are ranked according to their correctness. Facts that have been voted down a significant number of times are removed whereas facts that were consistently voted up are added to a set of approved facts that will be considered for adding to the LoD cloud. Explicit evaluation provides strong indications of the correctness of statements, but it requires active user involvement. At this point in the project, we have not collected enough validation data to give a meaningful evaluation. However, below is a selection of facts that have been approved (and scored) by domain scientists. Appendix A gives the full list of evaluated triples in Table A.1.

6.1.2 Validation in Use

Implicit (or extrinsic) evaluation is an evaluation technique in which the outcome of a computational task is not evaluated with respect to a set of standard goals, but by analyzing the impact of the outcome to a task. It has first been described by Bannon (1996) for applications in the design field. Recently, extrinsic evaluation has been widely applied in evaluating text summarization. but it’s applicability has also been challenged (Gillick and Liu, 2010).

Implicit validation is less straightforward in associating a measure of truth with a statement, but can still give very strong indications for the correctness of a fact, provided there is enough traffic on the site to get statistically meaningful results. Implicit validation has
therefore mostly been used on search- and classification results in recommender systems, for example for news retrieval (Das et al., 2007). Agichtein et al. (2006a,b) present an elaborate analysis of browsing behavior to build a feedback model. To account for differences in a model of user behavior when extracted facts are evaluated we added a few criteria. The following list shows implicit validation strategies for complete models as well as for single extracted facts. Given are expected user behavior in cases of correct or erroneous facts and potential reasons/remedies.

1. Verification of the model’s coverage:

   (a) Alternate between search results that were biased by the model and unbiased search results.

   → If many unbiased links are chosen, the model needs to be updated to reflect the user’s choices.

2. Verification of the extracted facts:

   (a) A fact is browsed very often by different users.

   → The fact is interesting to many users.

   → The fact is surprising and interesting, but may be incorrect.

   (b) A user leaves a page quickly after browsing to it via an extracted fact.

   → The information indicated by the triple was either not available or not recognized.

   → The triple was found interesting, but immediately recognized as incorrect.
(c) A user starts a new search and chooses a result that was a target page for a triple on the previous page.

→ The triple was not displayed prominently enough on the previous page.

→ The term denoting the subject of the triple was not the synonym the user expected.

→ The link through relationship and object of the triple was not obvious enough.

(d) A user follows a trail of multiple triples through a variety of documents.

→ The triples that were browsed have a high probability of being correct and support is added to the triples.

→ If the trail was longer than suggested by a small-world phenomenon, initial triples may have been incorrect, but led to interesting ones. For this reason, only the last \( k \) triples of the trail should garner support or the support should increase for the last \( k \) triples in the trail.

→ The last triple in the trail may have been incorrect and led to browsing results that caused the user to stop browsing. For this reason, the last triple of the trail should be treated with caution.

In order to find common behaviors for correct and for incorrect information, some triples that have already been identified as incorrect must be provided as candidates to the user as well as many that are known to be correct.
6.2 Discussion

6.2.1 User feedback to focused browsing

Qualitative evaluations of Scooner (Cameron et al., 2010; Kavuluru et al., 2012) were conducted by five researchers from the human effectiveness directorate of the AFRL. The search feature and the focused browsing Scooner were reportedly useful and convenient for document research tasks that go beyond regular PubMed search.

Two researchers reported that they were able to save significant time relative to their experience with PubMed. The head of the team reported that it saved him a lot of efforts in easily delegating tasks to his team members by sharing his sessions with his comments and notes on the various abstracts in the workbench. The evaluators noted that Scooner helped them stay focused in the cognitive research area, which is one of the original goals of the framework. They felt that the narrow focus helped them to perform in-depth exploration of specific topics without significant perusing of PubMed results. They also found the persistent projects and collaborative features made it easier to organize their search tasks. (Kavuluru et al. (2012))

6.2.2 Qualitative Evaluation of browsed facts

Due to the relative sparsity of browsing data from the semantic browser at this early stage of the application, only a small qualitative analysis of the search and browsing history can be performed. However, the users of the Scooner browser were all experts in the field of
cognitive neuroscience and hence ideal candidates for testing a dedicated domain model.

The preliminary analysis shows a promising trend.

Table 6.1: Fact browsing statistics. The table shows the assertion that was followed, the number of times it was accessed and its correctness according to domain experts.

<table>
<thead>
<tr>
<th>Assertion</th>
<th>#access</th>
<th>True/False/None</th>
</tr>
</thead>
<tbody>
<tr>
<td>alzheimer’s disease → arise → psen1 mutation</td>
<td>46</td>
<td>True</td>
</tr>
<tr>
<td>brain derived neurotroph factor bdnf → act →</td>
<td></td>
<td></td>
</tr>
<tr>
<td>length trkb receptor trkb fl</td>
<td>36</td>
<td>True</td>
</tr>
<tr>
<td>ps1 mutat → associated → famili alzheim disas</td>
<td>32</td>
<td>True</td>
</tr>
<tr>
<td>nootropic agent → related_entity → cognitive enhancement</td>
<td>32</td>
<td>True</td>
</tr>
<tr>
<td>vasoactive intestinal peptide (vip) → increase →</td>
<td></td>
<td></td>
</tr>
<tr>
<td>catecholamin biosynthesis</td>
<td>28</td>
<td>True</td>
</tr>
<tr>
<td>mutation of ps1 → causes → early onset of alzheimer’s disease</td>
<td>21</td>
<td>True</td>
</tr>
<tr>
<td>catecholamin → induces → beta adrenergic receptor activity</td>
<td>21</td>
<td>True</td>
</tr>
<tr>
<td>ps1 mutation → causes → ad pathology</td>
<td>18</td>
<td>True</td>
</tr>
<tr>
<td>nootropic agent → type → cognitive enhancer</td>
<td>16</td>
<td>True</td>
</tr>
<tr>
<td>ps1 mutat → causes → famili alzheim disas</td>
<td>15</td>
<td>True</td>
</tr>
<tr>
<td>beta adrenergic receptor → involved →</td>
<td></td>
<td></td>
</tr>
<tr>
<td>contextual fear conditioning</td>
<td>15</td>
<td>True</td>
</tr>
<tr>
<td>positive cost effect nootropes → related_category → nootropes</td>
<td>12</td>
<td>None</td>
</tr>
<tr>
<td>gene → involved_in → neurogenesis</td>
<td>11</td>
<td>None</td>
</tr>
<tr>
<td>nootropic drug → knows → human</td>
<td>9</td>
<td>False</td>
</tr>
<tr>
<td>muscarinic activ → facilitates → long term potentiation ltp</td>
<td>9</td>
<td>True</td>
</tr>
<tr>
<td>nootropic agent → related_entity → memory enhancing drug</td>
<td>8</td>
<td>True</td>
</tr>
<tr>
<td>nerve growth factor → inhibits → mitogen activ protein kinase</td>
<td>8</td>
<td>False</td>
</tr>
<tr>
<td>nootropic agent → related_entity → brain booster</td>
<td>7</td>
<td>True</td>
</tr>
<tr>
<td>long term potentiation ltp → induced_by → injury hippocampus</td>
<td>7</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>15 / 2 / 4</td>
<td></td>
</tr>
</tbody>
</table>

To give a qualitative analysis of Implicit Validation, I gathered all triples in the model that were browsed at least 7 times by the AFRL researchers and manually evaluated their correctness. Table 6.1 shows an analysis of these triples. Each row in the table represents an extracted statement that was used to find new literature pertaining to the content of the statement, the number of times the statement was accessed in a browsing-trail and its
factual correctness. Statements that are accessed often were generally true in this qualitative assessment. Some statements lacked a truth value or were trivial, such as those involving the related category relationship and were labeled none for “no truth value”, even if they were trivially correct. Due to the small sample of browsing data, no definite conclusion can be drawn. The users also mostly browsed short paths, so some of the strategies from the above list could not be examined. However, the results in the table give supporting evidence to the claim that users help identify correct statements just by using them in information exploration.

Furthermore, according to the interviews with the domain scientists, focused browsing was helpful in finding pertinent information and gave an incentive to interact with the extracted information in the domain model.

Even though implicit validation can give a good idea of which statements are valid, a final call has to be made using explicit validation. It is always possible that statements were accessed because they seemed interesting or surprising, which can be the case with false statements. However, the implicit validation strategy can weed out many incorrect, uninteresting and tautological statements.

6.3 Propagation of validated statements

Statements that have been sufficiently validated should be incorporated into the overall background knowledge. The top-down extraction that was chosen in this dissertation allows for a seamless integration of statements into knowledge bases that are based on Linked open Data concepts, because the concepts that are used for fact extraction are already part
of LoD. Alignment of concepts and relationship types is thus not an issue. When model extraction is done in a purely bottom-up fashion, for example by using NER techniques in conjunction with *Structural Targeting open IE* (Banko et al., 2008) to find previously unknown concepts, entities and relationship types, alignment becomes difficult. However, there is significant active research in ontology alignment that can ameliorate these problems; see for example (Noy, 2004) and (Kalfoglou and Schorlemmer, 2003) for detailed surveys on ontology alignment. In the context of this work we developed alignment techniques using Expectation Maximization (Doshi et al., 2009; Doshi and Thomas, 2006). Other methods that are well suited in the context of Web knowledge make use of Linked open Data and Wikipedia (Jain et al., 2010).

In the alignment phase it is once more beneficial that information is extracted in the form of focused models. Context helps disambiguate concept descriptions when they are extracted bottom-up (in the top-down case, ambiguity is already eliminated).

### 6.4 Conclusion

A combination of implicit and explicit validation can reduce the time that expert evaluators in dedicate solely to evaluating statements. This is particularly important, because the time spent on explicit evaluation is usually time lost for the domain expert. It can be assumed that the expert would not have been actively interested in many of the statements that are explicitly validated. So reducing the effort is of great importance.

However, it is likely that many experts (or users in general) will still have an interest in giving some of their time to validating knowledge. Developments in the context of Web 2.0...
give a good indication of peoples’ willingness to “help out” in many kinds of knowledge gathering tasks. Wikipedia is just one example of the dedication that is shown by users to create information in their area of expertise. In this work, both extraction and evaluation of knowledge focuses on particular fields that people are interested in and are willing to spend a little extra time to improve. The next chapter will give a general vision of evaluation and problem solving on the web.
Conclusion

This dissertation addressed the question of knowledge acquisition in a system from the points of view of Knowledge Engineering and automated Knowledge Extraction. The chapters explored techniques as well as opportunities and problems associated with either approach. Focus was thereby on automated ways of increasing the amount of formally available knowledge, albeit the extraction in the case of the Knowledge Engineering approach is highly specialized and makes use of the manually created T-Box knowledge more rigorously.

The Knowledge Engineering perspective should be taken when a domain needs to be deeply explored and the experts’ explicit as well as tacit knowledge needs to be formalized. Automated Extraction, on the other hand, is useful for more interactive tasks in which users utilize extracted domain models for IR tasks to explore new interests or gain more depth in fields of their expertise.

The Knowledge Engineering example showed how an in-depth representation of do-
main knowledge can aid a focused extraction of new knowledge. In that case, the extraction algorithm enriched the bare facts (or in the specific application molecular structures) that were extracted with the knowledge that had been formalized in the schema and in the archetypal instances. Thus, the extraction actually increased the information content of the extracted structures with knowledge that was previously tacit. In this restricted extraction setting it was possible to achieve 100% precision, because the possible compositions of the extracted structures were highly restricted and followed rules that were encoded in the T-Box and in the Archetypal Instances.

I also presented a framework for continuous knowledge acquisition in a system that pertain to the modeling of new knowledge, the extraction of focused information and the validation of this extracted information. This was done with a loosely connected system in mind where storage, production, extraction and validation of knowledge are conceptually, spatially and temporally separated. This mirrors the reality on the Web, where millions of users produce and consume knowledge whenever they please and whenever it is necessary.

The dissertation considered the acquisition of new facts using IE techniques aided by available background knowledge in the form of Linked open Data and a sufficient amount of information in mostly trusted sources, such as Wikipedia and MedLine. The formal knowledge contained in Linked open Data repositories is less structured and less focused than it is the case in ontologies such as GlycO. It is thus not possible to use this background knowledge to assert rules that can guide the extraction process in the same manner as in GlycO. However, being able to start an extraction on the basis of concept pairs rather than having to identify named entities in text ensures that the extracted statements refer to actual concepts. The results of the extraction show a precision of 79% for named relationships.
in a previously unknown scientific field of interest and medium to high precisions in non-scientific fields of inquiry. These results in previously unseen domains validated the results obtained by evaluating with respect to UMLS and DBPedia gold standards.

The precision of the IE algorithm was comparable to results in related research that were achieved using more involved NLP methods to pre-process the features used for classification rather than use unprocessed surface patterns as in our case. The low-recall problem that most pattern-based IE methods suffer from has been addressed with a pattern generalization technique that improves recall significantly without major loss of precision.

A further reason why recall still cannot fully measure up to NLP-based systems is also that NLP-based extraction usually operates on a per-sentence level and thus extracts statements that are found only once. Whereas this method is useful for tasks such as document summarization or extraction of case information from legal documents, it is more error-prone than the Macro-reading approach that is taken here. Macro-reading ensures that a statement has proper support from multiple mentions in the text corpus. This assures that the extracted statements are of actual ontological importance and not incidental mentions of two concepts in the same sentence.

A positive-only classifier makes the use of LoD training data possible without human intervention. Facing the problem of missing negative training data, a combination of entropy and pertinence computations was used to give pattern features more discriminative power when they are highly indicative of a particular relationship type on the one hand and to ensure that patterns do not compete when they express relationships with similar semantics on the other.

No matter how well an IE-based approach fares in terms of precision, the statements
that are extracted will not count as knowledge until they are validated by a justification procedure. The approaches to justification that were taken in this dissertation are epistemologically based on coherence and consensus. A coherence-based justification procedure can be in the form of a deductive rule-based system, given the extracted statements can be entailed by a given knowledge base, as is the case in the extraction for the GlycO ontology. Consensus is seen here as an agreement between human agents on the validity and truthfulness of a statement. The methods that were described in this dissertation for consensus-based validation included explicit and implicit validation/evaluation of statements.

*Explicit validation* describes the traditional method in which a group of experts agrees on the truth or falsity of a statement. This is thus far the only way to make a final assessment on the validity of a statement. However, it was shown how an *implicit validation* (or evaluation in use) can be used to narrow the list of statements down to a small number of statements that are likely to be true. The hypothesis is that user behavior can give an indication of the correctness of a statement. Statements that are used more often than others are more likely to be correct. The ’use’ of a statement can thereby take many forms. In this case statements were used as guidelines for exploratory browsing. The “Evaluation in Use” method was tested using the semantic browser *Scooner* (Cameron et al., 2010; Kavuluru et al., 2012). The results are promising, even though the amount of data was insufficient to make statistically significant claims about an overall success.

The hypothesis of this dissertation is that the overall knowledge in a system can be increased by focused acquisition of domain and task knowledge, its validation and subsequent addition to the previously available knowledge. The automatic model creation part of this work succeeds at this task not only by having a high-quality extraction system, but
by providing means to verify the extracted assertions in use. Information retrieval tasks account for a large amount of Web activity. Allowing formal context models to be incorporated into these tasks can achieve a better coverage of the domain of interest. In turn, the user gives information back to the system while using the model. The evaluation of user behavior on a scientific IR task in chapter 6 shows that the users intuitively used correct assertions more often than incorrect ones to explore new information.

One of the corollaries of the hypothesis is that evaluation is an intrinsic part of knowledge extraction, rather than an afterthought to verify algorithmic performance. Once evaluation is seen as such and once it is made accessible to a general user base via different means of social computation, it is straightforward to integrate it into a knowledge acquisition process.

7.1 Outlook

This work was mostly concerned with acquiring knowledge. In order to make knowledge acquisition in a system sustainable, knowledge management needs to assure that the system can maintain a state in which the beliefs that are held are always justified. This includes finding systematic ways of holding contradictory beliefs and resolving contradictions when the evidence is overwhelmingly in favor of a specific view.

If, as it is likely to happen, statements that are more correct than others will prevail in the long run, the collections of background knowledge should adapt and replace incorrect assertions with correct ones. The crucial task for future work is to create systems that can detect and cope with inconsistencies. Systems that treat evidence that contradicts their
beliefs with caution, but are able to incorporate the inconsistencies and maybe eventually shift to a new set of predominant beliefs, as described by Kuhn (1996) in *The structure of scientific revolutions* for the example of scientific communities. In Kuhn’s view, the conservatism of the scientific community when it comes to adopting new beliefs is a necessary part of the process, because it stabilizes the belief system, but eventually better or more coherent ideas will get the upper hand. Knowledge management systems on the web will have to be cautious as well, but also need to allow for these changes to happen.

Ideally, knowledge bases would allow for the storage of multiple beliefs and belief systems. The CYC KB, for example, has multiple so-called microtheories that are individually consistent, but can contain facts that are inconsistent with facts from other microtheories. This allows to have e.g. a formal representation of both Newtonian physics and of quantum physics, both of which can be applied in different scenarios. However, even CYC assumes that there is one correct set of facts per microtheory, not multiple competing sets.

Knowledge management systems on the Web should ideally store facts associated with a degree of belief and a theory that the fact is assumed to be valid in. A scientific community needs to be constantly vigilant that the knowledge on which grounds it is operating is still valid. In the face of massive amounts of knowledge it is important that the system helps verifying the stored knowledge. Moreover, even though newer and maybe contradictory facts may come in, these need not necessarily replace the old ones as a predominant theory within the system.

URL http://portal.acm.org/citation.cfm?doid=1148170.1148177

URL http://portal.acm.org/citation.cfm?id=1148175&amp;dl=GUIDE,


American 284 (5), 34–43.

URL http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21


URL http://dl.acm.org/citation.cfm?id=1367497.1367561


URL http://citeseer.ist.psu.edu/695326.html


URL http://dx.doi.org/http://doi.acm.org/10.1145/505168.505197

information science, 117–151.

URL http://books.google.com/books?hl=en&lr=&id=_0sDF-0vT0wC&oi=fnd&pg=PA124&dq=A+Theory+of+Granular+Partitions&ots=uLM423xNp3&sig=1HADnPeFwFTR37fB266qRGbhW38


URL http://www.igi-global.com/articles/details.asp?ID=35386


URL http://linkinghub.elsevier.com/retrieve/pii/S1570826809000225


Bohne-Lang, A., Lang, E., Förster, T., von der Lieth, C. W., 2001. LINUCS: Linear Nota-
tion for Unique Description of Carbohydrate Sequences. Carbohydrate research 336 (1), 1–11.


URL http://doc.utwente.nl/17864/


URL http://books.google.com/books?hl=en&amp;lr=&amp;id=gikMilZMFMMC&amp;oi=fnd&amp;pg=PA3&amp;dq=Ontology+Learning+from+Text++An+Overview&amp;ots=QqU3dfSfbS&amp;sig=SLQ6K1u-IUPRR9qK41XyFO18XeU
Cameron, D., Mendes, P. N., Sheth, A. P., 2010. Semantics-Empowered Text Exploration for Knowledge Discovery. ACMSE Conference.


URL http://portal.acm.org/citation.cfm?id=1718501&dl=GUIDE,


URL http://www.springerlink.com/index/ldb45u6uqgeyr7pq.pdf

URL http://portal.acm.org/citation.cfm?doid=1631850.1631854

URL http://portal.acm.org/citation.cfm?id=1242572.1242610


URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.961


URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.1022&rep=rep1&type=pdf


URL http://dl.acm.org/citation.cfm?id=1866718


URL

URL http://labs.google.com/sets


URL http://portal.acm.org/citation.cfm?id=503150&dl=GUIDE,


URL http://portal.acm.org/citation.cfm?id=79161


URL http://dx.doi.org/10.1002/(SICI)1098-237X(200005)84:3<352::AID-SCE3>3.0.CO;2-J

URL http://www.rci.rutgers.edu/~jasoncs/knowledgeactionfinal.pdf


URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4305569


Horrocks, I., Patel-Schneider, P. F., Harmelen, F. V., Dec. 2003. From SHIQ and RDF to

URL http://linkinghub.elsevier.com/retrieve/pii/S1570826803000027


URL http://linkinghub.elsevier.com/retrieve/pii/S1570826812000121


URL http://www.springerlink.com/index/Y16058375754U221.pdf


URL http://portal.acm.org/citation.cfm?id=775126


URL http://analysis.oxfordjournals.org/content/36/3/150.full.pdf


URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=920602


URL http://portal.acm.org/citation.cfm?id=1690287&dl=GUIDE,


Navigli, R., Velardi, P., Faralli, S., 2011. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. ... of the Twenty-Second international joint ... , 1872–1877.

URL http://dl.acm.org/citation.cfm?id=2283715


URL http://portal.acm.org/citation.cfm?doid=360018.360022


Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A., 2006. Organizing and Searching the


URL http://www.cse.ohio-state.edu/~ruan/papers/some11.pdf


URL http://www.springerlink.com/index/kh10n04667j34w04.pdf


URL http://content.apa.org/books/11974-000


URL http://www.jstor.org/stable/2248381


URL http://iospress.metapress.com/index/x430q0km475q5577.pdf


URL http://portal.acm.org/citation.cfm?id=234209


URL http://dx.doi.org/10.1109/MIC.2002.1020330


URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.3649&rep=rep1&type=pdf


URL http://books.google.com/books?hl=en&lr=&id=MkgR0zyTba4C&oi=fnd&pg=PA13&ots=Dlvk_9tEXW&sig=wtT0-hmDCmHnv5viTnkTHzieiQQ

URL http://acl.ldc.upenn.edu/P/P06/P06-1040.pdf

URL http://arxiv.org/pdf/1003.1141


URL http://biomed.uninet.edu/2011/n2/velasco.html


URL http://portal.acm.org/citation.cfm?id=253801


URL http://portal.acm.org/citation.cfm?doid=1519103.1519113


URL http://www.springerlink.com/index/M811790058246821.pdf


URL http://www.garlic.com/~lynn/


Yu, X., King, I., Lyu, M., 2011. Towards a Top-Down and Bottom-Up Bidirectional Approach to Joint Information Extraction. ... 20th ACM international conference on Infor-
Zins, C., 2006. Redefining Information Science: from information Science to Knowledge
URL http://www.emeraldinsight.com/10.1108/00220410610673846

Zirn, C., Nastase, V., Strube, M., 2008. Distinguishing Between Instances and Classes in
Appendix A
Table A.1: All triples that were evaluated by neuroscientists at the Wright Patterson Air
Force Research lab. The first column gives the expert score with 1-2 being incorrect, 3-4
correct, but trivial, 5-6 correct general information and 7-9 being correct and not commonly
known

<table>
<thead>
<tr>
<th>Expert Score</th>
<th>Triple</th>
<th>Confidence Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Fmr1 → gene, plays, role_in, process → Synaptic plasticity</td>
<td>0.819697312</td>
</tr>
<tr>
<td>9</td>
<td>Fmr1 → gene_product, plays, role_in, biological_process → Synaptic plasticity</td>
<td>0.793415599</td>
</tr>
<tr>
<td>9</td>
<td>Pallidotomy → has, direct, procedure, site → Globus pallidus</td>
<td>0.90660523</td>
</tr>
<tr>
<td>9</td>
<td>Commissural_Fibers → has, physical, part_of, anatomic, structure → Posterior commissure</td>
<td>0.866879767</td>
</tr>
<tr>
<td>9</td>
<td>Limbic_System → has, regional, part → Parahippocampal gyrus</td>
<td>0.968839511</td>
</tr>
<tr>
<td>9</td>
<td>Diencephalon → has, regional, part → Epithalamus</td>
<td>0.958695789</td>
</tr>
<tr>
<td>9</td>
<td>Auditory_System → has, regional, part → Superior temporal gyrus</td>
<td>0.910353812</td>
</tr>
<tr>
<td>9</td>
<td>Diencephalon → has, regional, part → Thalamus</td>
<td>0.893053594</td>
</tr>
<tr>
<td>9</td>
<td>Incus → part_of → Middle ear</td>
<td>0.893174494</td>
</tr>
<tr>
<td>9</td>
<td>Cerebellum → receives, input_from → Solitary nucleus</td>
<td>0.985406764</td>
</tr>
<tr>
<td>9</td>
<td>Cochlear_Nuclei → receives, input_from → Pontine nuclei</td>
<td>0.958361221</td>
</tr>
<tr>
<td>9</td>
<td>Zona_Incerta → receives, input_from → Cingulate cortex</td>
<td>0.921019431</td>
</tr>
<tr>
<td>9</td>
<td>Nucleus_Raphes_Magnus → receives, input_from → Periaqueductal gray</td>
<td>0.916757239</td>
</tr>
<tr>
<td>9</td>
<td>Pineal_Gland → has, regional, part_of → Epithalamus</td>
<td>0.922776159</td>
</tr>
<tr>
<td>9</td>
<td>Laterodorsal_Tegmental_Nucleus → sends, output_to → Hypothalamus</td>
<td>0.891744347</td>
</tr>
<tr>
<td>9</td>
<td>Dentate_Nucleus → sends, output_to → Thalamus</td>
<td>0.874008013</td>
</tr>
<tr>
<td>8</td>
<td>Nucleus_Ambiguus → anatomic, structure, is, physical, part_of → Medulla oblongata</td>
<td>0.752122095</td>
</tr>
<tr>
<td>8</td>
<td>Pedunculopontine_Nucleus → anatomic, structure, is, physical, part_of → Reticular activating system</td>
<td>0.747968809</td>
</tr>
<tr>
<td>8</td>
<td>Mitral_Cell → anatomic, structure, is, physical, part_of → Olfactory bulb</td>
<td>0.733611609</td>
</tr>
<tr>
<td>8</td>
<td>Aniracetam → associated_with → Anxiolytic</td>
<td>0.710746638</td>
</tr>
<tr>
<td>8</td>
<td>Oligodendroglioma → disease, has, associated, anatomic, site → Frontal lobe</td>
<td>0.733586061</td>
</tr>
<tr>
<td>8</td>
<td>Transmissible_Spongiform_Encephalopathy → disease, has, primary, anatomic, site → Brain</td>
<td>0.888095433</td>
</tr>
<tr>
<td>8</td>
<td>Hippocampus → finding, site_of → Autobiographical memory</td>
<td>0.841378113</td>
</tr>
<tr>
<td>8</td>
<td>Expressive_Aphasia → has, associated, morphology → Aphasia</td>
<td>0.877299231</td>
</tr>
<tr>
<td>8</td>
<td>Subdural_Hematoma → has, finding, site → Meninges</td>
<td>0.834612736</td>
</tr>
<tr>
<td>8</td>
<td>Neuroacanthocytosis → has, manifestation → Basal ganglia</td>
<td>0.780057522</td>
</tr>
<tr>
<td>8</td>
<td>Nervous_System → has, part → Autonomic nervous system</td>
<td>0.883629703</td>
</tr>
<tr>
<td>8</td>
<td>Middle_Ear → has, part → Incus</td>
<td>0.869556441</td>
</tr>
<tr>
<td>8</td>
<td>Auditory_System → has, part → Medial geniculate nucleus</td>
<td>0.85270525</td>
</tr>
<tr>
<td>8</td>
<td>Limbic_System → has, part → Parahippocampal gyrus</td>
<td>0.839786941</td>
</tr>
<tr>
<td>8</td>
<td>Auditory_System → has, part → Dorsal cochlear nucleus</td>
<td>0.740796576</td>
</tr>
<tr>
<td>8</td>
<td>Mesencephalon → has, physical, part_of, anatomic, structure → Corpora quadrigemina</td>
<td>0.900254672</td>
</tr>
<tr>
<td>8</td>
<td>Auditory_System → has, physical, part_of, anatomic, structure → Superior temporal gyrus</td>
<td>0.70051603</td>
</tr>
<tr>
<td>8</td>
<td>Limbic_System → has, regional, part → Amygdala</td>
<td>0.891848602</td>
</tr>
<tr>
<td>8</td>
<td>Diencephalon → has, regional, part → Subthalamus</td>
<td>0.883657763</td>
</tr>
<tr>
<td>8</td>
<td>Parietal_Lobe → has, regional, part → Postcentral gyrus</td>
<td>0.864660282</td>
</tr>
<tr>
<td>8</td>
<td>Limbic_System → has, regional, part → Cingulate gyrus</td>
<td>0.850810471</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Meninges → is associated_anatomic_site_of → Meningioma</td>
<td>0.858226767</td>
</tr>
<tr>
<td>8</td>
<td>Hippocampus → location_of → Autobiographical memory</td>
<td>0.840714332</td>
</tr>
<tr>
<td>8</td>
<td>Autonomic_Nervous_System → part_of → Nervous system</td>
<td>0.884414896</td>
</tr>
<tr>
<td>8</td>
<td>Thalamic_Reticular_Nucleus → receives_input_from → Cerebral cortex</td>
<td>0.894396768</td>
</tr>
<tr>
<td>8</td>
<td>Subthalamic_Nucleus → receives_input_from → Globus pallidus</td>
<td>0.879425604</td>
</tr>
<tr>
<td>8</td>
<td>Putamen → receives_input_from → Cerebral cortex</td>
<td>0.84230226</td>
</tr>
<tr>
<td>8</td>
<td>Primary_Auditory_Cortex → receives_input_from → Medial geniculate nucleus</td>
<td>0.830179738</td>
</tr>
<tr>
<td>8</td>
<td>Septal_Nuclei → receives_input_from → Amygdala</td>
<td>0.746130292</td>
</tr>
<tr>
<td>8</td>
<td>Ventral_Lateral_Nucleus → receives_input_from → Cerebellum</td>
<td>0.719892779</td>
</tr>
<tr>
<td>8</td>
<td>Superior_Colliculus → receives_input_from → Spinal trigeminal nucleus</td>
<td>0.719472228</td>
</tr>
<tr>
<td>8</td>
<td>Sensory_Cortex → regional_part_of → Cerebral cortex</td>
<td>0.961294727</td>
</tr>
<tr>
<td>8</td>
<td>Amygdala → regional_part_of → Limbic system</td>
<td>0.88022019</td>
</tr>
<tr>
<td>8</td>
<td>Hypothalamus → regional_part_of → Diencephalon</td>
<td>0.877903089</td>
</tr>
<tr>
<td>8</td>
<td>Cingulate_Gyrus → regional_part_of → Limbic system</td>
<td>0.867653586</td>
</tr>
<tr>
<td>8</td>
<td>Habenular_Nuclei → regional_part_of → Epithalamus</td>
<td>0.858817211</td>
</tr>
<tr>
<td>8</td>
<td>Cuneus → regional_part_of → Occipital lobe</td>
<td>0.841259987</td>
</tr>
<tr>
<td>8</td>
<td>Area_Postrema → regional_part_of → Medulla oblongata</td>
<td>0.820858496</td>
</tr>
<tr>
<td>8</td>
<td>Superior_Frontal_Gyrus → regional_part_of → Frontal lobe</td>
<td>0.79378262</td>
</tr>
<tr>
<td>8</td>
<td>Dorsal_Cochlear_Nucleus → sends_output_to → Inferior colliculus</td>
<td>0.882634976</td>
</tr>
<tr>
<td>8</td>
<td>Laterodorsal_Tegmental_Nucleus → sends_output_to → Thalamus</td>
<td>0.772257799</td>
</tr>
<tr>
<td>8</td>
<td>Cuneate_Nucleus → sends_output_to → Thalamus</td>
<td>0.754484605</td>
</tr>
<tr>
<td>8</td>
<td>Subthalamus → sends_output_to → Caudate nucleus</td>
<td>0.754484605</td>
</tr>
<tr>
<td>8</td>
<td>Reticular_Formation → sib_in_branch_of → Reticular activating system</td>
<td>0.890530287</td>
</tr>
<tr>
<td>7</td>
<td>Iconic_Memory → subtype_of → Sensory memory</td>
<td>0.751030417</td>
</tr>
<tr>
<td>7</td>
<td>Epithalamus → anatomic_structure_has_location → Limbic system</td>
<td>0.791208877</td>
</tr>
<tr>
<td>7</td>
<td>Ependyma → anatomic_structure_is_physical_part_of → Central nervous system</td>
<td>0.70936316</td>
</tr>
<tr>
<td>7</td>
<td>Communication_Disorder → associated_morphology_of → Stuttering</td>
<td>0.739667134</td>
</tr>
<tr>
<td>7</td>
<td>Stuttering → associated_with → Fear</td>
<td>0.970428842</td>
</tr>
<tr>
<td>7</td>
<td>Traumatic_Brain_Injury → associated_with → Panic disorder</td>
<td>0.922804044</td>
</tr>
<tr>
<td>7</td>
<td>Awareness → biological_process_is_part_of_process → Cognition</td>
<td>0.819724512</td>
</tr>
<tr>
<td>7</td>
<td>Tourette_Syndrome → direct_procedure_site_of → Deep brain stimulation</td>
<td>0.744801243</td>
</tr>
<tr>
<td>7</td>
<td>Language → direct_procedure_site_of → Angular gyrus</td>
<td>0.737841947</td>
</tr>
<tr>
<td>7</td>
<td>Neuroglycopenia → disease_has_associated_anatomic_site → Central nervous system</td>
<td>0.794373083</td>
</tr>
<tr>
<td>7</td>
<td>Putamen → finding_site_of → Schizophrenia</td>
<td>0.747974699</td>
</tr>
<tr>
<td>7</td>
<td>Magnetoencephalography → finding_site_of → Schizophrenia</td>
<td>0.737540472</td>
</tr>
<tr>
<td>7</td>
<td>Visual_Cortex → has_component → Consciousness</td>
<td>0.875193585</td>
</tr>
<tr>
<td>7</td>
<td>Mesencephalon → has_constitutional_part → Cerebral aqueduc</td>
<td>0.837265201</td>
</tr>
<tr>
<td>7</td>
<td>Systematic_Desensitization → has_direct_procedure_site → Anxiety</td>
<td>0.727521128</td>
</tr>
<tr>
<td>7</td>
<td>Spinal_Cord → has_part → Spinthalamic tract</td>
<td>0.91422796</td>
</tr>
<tr>
<td>7</td>
<td>Middle_Ear → has_part → Malleus</td>
<td>0.849243005</td>
</tr>
<tr>
<td>7</td>
<td>Diencephalon → has_part → Epithalamus</td>
<td>0.84567423</td>
</tr>
<tr>
<td>7</td>
<td>Neocortex → has_part → Occipital lobe</td>
<td>0.816005253</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Limbic_System → has_part → Amygdala</td>
<td>0.809997106</td>
</tr>
<tr>
<td>7</td>
<td>Middle_Ear → has_part → Stapes</td>
<td>0.805686408</td>
</tr>
<tr>
<td>7</td>
<td>Diencephalon → has_part → Thalamus</td>
<td>0.781319654</td>
</tr>
<tr>
<td>7</td>
<td>Limbic_System → has_part → Hippocampus</td>
<td>0.765838266</td>
</tr>
<tr>
<td>7</td>
<td>Frontal_Lobe → has_part → Orbitofrontal cortex</td>
<td>0.722890224</td>
</tr>
<tr>
<td>7</td>
<td>Cochlea → has_part → Scala tympani</td>
<td>0.715357545</td>
</tr>
<tr>
<td>7</td>
<td>Frontal_Lobe → has_physical_part_of_anatomic_structure → Middle frontal gyrus</td>
<td>0.839904134</td>
</tr>
<tr>
<td>7</td>
<td>Auditory_System → has_physical_part_of_anatomic_structure → Dorsal cochlear nucleus</td>
<td>0.811408884</td>
</tr>
<tr>
<td>7</td>
<td>Auditory_System → has_physical_part_of_anatomic_structure → Middle ear</td>
<td>0.759384836</td>
</tr>
<tr>
<td>7</td>
<td>Mesencephalon → has_physical_part_of_anatomic_structure → Substantia nigra</td>
<td>0.758960896</td>
</tr>
<tr>
<td>7</td>
<td>Neocortex → has_physical_part_of_anatomic_structure → Occipital lobe</td>
<td>0.736458519</td>
</tr>
<tr>
<td>7</td>
<td>Parietal_Lobe → has_physical_part_of_anatomic_structure → Intraparietal sulcus</td>
<td>0.718529417</td>
</tr>
<tr>
<td>7</td>
<td>Olfactory_Bulb → has_physical_part_of_anatomic_structure → Mitral cell</td>
<td>0.716432379</td>
</tr>
<tr>
<td>7</td>
<td>Grey_Matter → has_regional_part → Nucleus accumbens</td>
<td>0.839378035</td>
</tr>
<tr>
<td>7</td>
<td>Occipital_Lobe → has_regional_part → Cuneus</td>
<td>0.830027624</td>
</tr>
<tr>
<td>7</td>
<td>Medulla_Oblongata → has_regional_part → Area postrema</td>
<td>0.805556787</td>
</tr>
<tr>
<td>7</td>
<td>Cerebral_Cortex → has_regional_part → Occipital lobe</td>
<td>0.802527817</td>
</tr>
<tr>
<td>7</td>
<td>Frontal_Lobe → has_regional_part → Superior frontal gyrus</td>
<td>0.801527769</td>
</tr>
<tr>
<td>7</td>
<td>Limbic_System → has_regional_part → Hippocampus</td>
<td>0.760541446</td>
</tr>
<tr>
<td>7</td>
<td>Limbic_System → has_regional_part → Entorhinal cortex</td>
<td>0.711848348</td>
</tr>
<tr>
<td>7</td>
<td>Frontal_Lobe → has_regional_part → Orbitofrontal cortex</td>
<td>0.702761995</td>
</tr>
<tr>
<td>7</td>
<td>Central_Nervous_System_Infection → is-associated_anatomic_site_of → Meningoencephalitis</td>
<td>0.905684103</td>
</tr>
<tr>
<td>7</td>
<td>Nervous_System → is_location_of_biological_process → Olfaction</td>
<td>0.864032896</td>
</tr>
<tr>
<td>7</td>
<td>Meninges → location_of → Meningioma</td>
<td>0.845212354</td>
</tr>
<tr>
<td>7</td>
<td>Putamen → location_of → Schizophrenia</td>
<td>0.700343897</td>
</tr>
<tr>
<td>7</td>
<td>Optic_Neuropathy → manifestation_of → Occipital lobe</td>
<td>0.781676079</td>
</tr>
<tr>
<td>7</td>
<td>Malleus → part_of → Middle ear</td>
<td>0.855206394</td>
</tr>
<tr>
<td>7</td>
<td>Stapes → part_of → Middle ear</td>
<td>0.82158611</td>
</tr>
<tr>
<td>7</td>
<td>Hypothalamus → part_of → Diencephalon</td>
<td>0.781857898</td>
</tr>
<tr>
<td>7</td>
<td>Amygdala → part_of → Diencephalon</td>
<td>0.777067624</td>
</tr>
<tr>
<td>7</td>
<td>Orbitofrontal_Cortex → part_of → Frontal lobe</td>
<td>0.742125782</td>
</tr>
<tr>
<td>7</td>
<td>Hippocampus → part_of → Limbic system</td>
<td>0.735494528</td>
</tr>
<tr>
<td>7</td>
<td>Tuber_Cinereum → part_of → Hypothalamus</td>
<td>0.729359174</td>
</tr>
<tr>
<td>7</td>
<td>Corpus_Striatum → receives_input_from → Subthalamic nucleus</td>
<td>0.815127151</td>
</tr>
<tr>
<td>7</td>
<td>Subthalamic_Nucleus → receives_input_from → Cerebral cortex</td>
<td>0.814815346</td>
</tr>
<tr>
<td>7</td>
<td>Third_Ventricle → regional_part_of → Diencephalon</td>
<td>0.794475384</td>
</tr>
<tr>
<td>7</td>
<td>Zona_Incerta → regional_part_of → Diencephalon</td>
<td>0.784311903</td>
</tr>
<tr>
<td>7</td>
<td>Inferior_Frontal_Gyrus → regional_part_of → Frontal lobe</td>
<td>0.736564185</td>
</tr>
<tr>
<td>7</td>
<td>Vestibular_Nuclei → regional_part_of → Medulla oblongata</td>
<td>0.726114984</td>
</tr>
<tr>
<td>7</td>
<td>Zona_Incerta → regional_part_of → Subthalamic</td>
<td>0.723404646</td>
</tr>
<tr>
<td>7</td>
<td>Glia_Limitans → regional_part_of → Central nervous system</td>
<td>0.706084792</td>
</tr>
<tr>
<td>7</td>
<td>Optic_Neuropathy → related_to → Occipital lobe</td>
<td>0.821458743</td>
</tr>
</tbody>
</table>
Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Primary_Motor_Cortex (\rightarrow) sends_output_to (\rightarrow) Spinal cord</td>
<td>0.766797877</td>
</tr>
<tr>
<td>7</td>
<td>Oculomotor_Nucleus (\rightarrow) sends_output_to (\rightarrow) Substantia nigra</td>
<td>0.711468614</td>
</tr>
<tr>
<td>7</td>
<td>Encephalization (\rightarrow) used_for (\rightarrow) Intelligence</td>
<td>0.754862729</td>
</tr>
<tr>
<td>6</td>
<td>Dorsal_Root_Ganglion (\rightarrow) anatomic_structure_is_physical_part_of (\rightarrow) Peripheral nervous system</td>
<td>0.905098255</td>
</tr>
<tr>
<td>6</td>
<td>Central_Sulcus (\rightarrow) anatomic_structure_is_physical_part_of (\rightarrow) Cerebral cortex</td>
<td>0.712249975</td>
</tr>
<tr>
<td>6</td>
<td>Neuropsychological_Test (\rightarrow) associated_with (\rightarrow) Intelligence_quotient</td>
<td>0.884035992</td>
</tr>
<tr>
<td>6</td>
<td>Artificial_Intelligence (\rightarrow) associated_with (\rightarrow) Cognitive_architecture</td>
<td>0.861247725</td>
</tr>
<tr>
<td>6</td>
<td>Nootropic (\rightarrow) associated_with (\rightarrow) Pyritinol</td>
<td>0.746529203</td>
</tr>
<tr>
<td>6</td>
<td>Alexithymia (\rightarrow) associated_with (\rightarrow) Feeling</td>
<td>0.727964183</td>
</tr>
<tr>
<td>6</td>
<td>Superior_Frontal_Gyrus (\rightarrow) attributed_continuous_with (\rightarrow) Middle frontal gyrus</td>
<td>0.88693016</td>
</tr>
<tr>
<td>6</td>
<td>Median_Eminence (\rightarrow) attributed_continuous_with (\rightarrow) Optic chiasm</td>
<td>0.71941239</td>
</tr>
<tr>
<td>6</td>
<td>Depersonalization_Disturbance (\rightarrow) classified_as (\rightarrow) Depersonalization</td>
<td>0.895507754</td>
</tr>
<tr>
<td>6</td>
<td>Affective_Spectrum (\rightarrow) definitional_manifestation_of (\rightarrow) Personality disorder</td>
<td>0.845612536</td>
</tr>
<tr>
<td>6</td>
<td>Traumatic_Brain_Injury (\rightarrow) disease (\rightarrow) has_associated_anatomic_site (\rightarrow) Central nervous system</td>
<td>0.749252905</td>
</tr>
<tr>
<td>6</td>
<td>Feline_Spongiform_Encephalopathy (\rightarrow) disease (\rightarrow) has_primary_anatomic_site (\rightarrow) Brain</td>
<td>0.967839537</td>
</tr>
<tr>
<td></td>
<td>Outer_Ear (\rightarrow) hasconstitutional_part (\rightarrow) Ear canal</td>
<td>0.743978374</td>
</tr>
<tr>
<td>6</td>
<td>Deep_Brain_Stimulation (\rightarrow) has_direct_procedure_site (\rightarrow) Tourette syndrome</td>
<td>0.728548232</td>
</tr>
<tr>
<td>6</td>
<td>Deep_Brain_Stimulation (\rightarrow) has_direct_procedure_site (\rightarrow) Epilepsy</td>
<td>0.710519656</td>
</tr>
<tr>
<td>6</td>
<td>Anterograde_Amnesia (\rightarrow) has_location (\rightarrow) Limbic system</td>
<td>0.744038302</td>
</tr>
<tr>
<td>6</td>
<td>Upper_Motor_Neuron (\rightarrow) has_location (\rightarrow) Spinal cord</td>
<td>0.70393282</td>
</tr>
<tr>
<td>6</td>
<td>Frontotemporal_Dementia (\rightarrow) has_location (\rightarrow) Disinhibition</td>
<td>0.893020727</td>
</tr>
<tr>
<td>6</td>
<td>Commissural_Fibers (\rightarrow) has_location (\rightarrow) Posterior commissure</td>
<td>0.800136749</td>
</tr>
<tr>
<td>6</td>
<td>Anterograde_Amnesia (\rightarrow) has_location (\rightarrow) Limbic system</td>
<td>0.746214097</td>
</tr>
<tr>
<td>6</td>
<td>Mania (\rightarrow) has_location (\rightarrow) Amygdala</td>
<td>0.704063875</td>
</tr>
<tr>
<td>6</td>
<td>Nervous_System (\rightarrow) has_part (\rightarrow) Somatic nervous system</td>
<td>0.95083267</td>
</tr>
<tr>
<td>6</td>
<td>Spinal_Cord (\rightarrow) has_part (\rightarrow) Rubrospinal tract</td>
<td>0.887127685</td>
</tr>
<tr>
<td>6</td>
<td>Neural_Pathway (\rightarrow) has_part (\rightarrow) Spinal cord</td>
<td>0.763176943</td>
</tr>
<tr>
<td>6</td>
<td>Auditory_System (\rightarrow) has_part (\rightarrow) Middle ear</td>
<td>0.719406566</td>
</tr>
<tr>
<td>6</td>
<td>Cognitive_Test (\rightarrow) has_part (\rightarrow) Episodic memory</td>
<td>0.701661166</td>
</tr>
<tr>
<td>6</td>
<td>Subarachnoid_Space (\rightarrow) has_physical_part_of_anatomic_structure (\rightarrow) Gyrus</td>
<td>0.904447659</td>
</tr>
<tr>
<td>6</td>
<td>Human_Brain (\rightarrow) has_physical_part_of_anatomic_structure (\rightarrow) Cerebral hemisphere</td>
<td>0.766786195</td>
</tr>
<tr>
<td>6</td>
<td>Central_Nervous_System (\rightarrow) has_physical_part_of_anatomic_structure (\rightarrow) Meninges</td>
<td>0.726389457</td>
</tr>
<tr>
<td>6</td>
<td>Frontal_Lobe (\rightarrow) has_regional_part (\rightarrow) Inferior frontal gyrus</td>
<td>0.760858604</td>
</tr>
<tr>
<td>6</td>
<td>Hypnagogia (\rightarrow) interprets (\rightarrow) Sleep</td>
<td>0.73841048</td>
</tr>
<tr>
<td>6</td>
<td>Barrel_Cortex (\rightarrow) part_of (\rightarrow) Neocortex</td>
<td>0.706503966</td>
</tr>
<tr>
<td>6</td>
<td>Inferior_Colliculus (\rightarrow) receives_input_from (\rightarrow) Lateral geniculate nucleus</td>
<td>0.710536998</td>
</tr>
<tr>
<td>5</td>
<td>Subthalamic_Nucleus (\rightarrow) receives_input_from (\rightarrow) Substantia nigra</td>
<td>0.745845269</td>
</tr>
<tr>
<td>6</td>
<td>Hippocampus (\rightarrow) regional_part_of (\rightarrow) Limbic system</td>
<td>0.767751532</td>
</tr>
<tr>
<td>6</td>
<td>Orbitofrontal_Cortex (\rightarrow) regional_part_of (\rightarrow) Frontal lobe</td>
<td>0.72359837</td>
</tr>
<tr>
<td>6</td>
<td>Psychosis (\rightarrow) related_to (\rightarrow) Cyclothymia</td>
<td>0.907952833</td>
</tr>
<tr>
<td>6</td>
<td>Intraoperative_Monitoring (\rightarrow) related_to (\rightarrow) Electroencephalography</td>
<td>0.735399009</td>
</tr>
<tr>
<td>6</td>
<td>Affective_Spectrum (\rightarrow) related_to (\rightarrow) Personality disorder</td>
<td>0.715460258</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Inferior_Cerebellar_Peduncle → sends_output_to → Spinocerebellar tract</td>
<td>0.750876553</td>
</tr>
<tr>
<td>6</td>
<td>Trichotillomania → use → Obsessive Compulsive Disorder</td>
<td>0.894323054</td>
</tr>
<tr>
<td>5</td>
<td>Autosomal_Dominant_Nocturnal_Frontal_Lobe_Epilepsy → anatomic_structure_is_physical_part_of → Frontal lobe</td>
<td>0.734194534</td>
</tr>
<tr>
<td>5</td>
<td>Ganglion_Cell → anatomic_structure_is_physical_part_of → Ganglion</td>
<td>0.712986875</td>
</tr>
<tr>
<td>5</td>
<td>Mitral_Cell → anatomic_structure_is_physical_part_of → Brain</td>
<td>0.708093321</td>
</tr>
<tr>
<td>1</td>
<td>Uncus → anatomic_structure_is_physical_part_of → Temporal lobe</td>
<td>0.734550904</td>
</tr>
<tr>
<td>5</td>
<td>Tourette_Syndrome → associated_with → Developmental disorder</td>
<td>0.846178083</td>
</tr>
<tr>
<td>5</td>
<td>Agoraphobia → associated_with → Separation anxiety disorder</td>
<td>0.796809775</td>
</tr>
<tr>
<td>5</td>
<td>Semantic_Memory → associated_with → Neural network</td>
<td>0.785453988</td>
</tr>
<tr>
<td>5</td>
<td>Human_Brain → associated_with → Brodmann area</td>
<td>0.76182265</td>
</tr>
<tr>
<td>5</td>
<td>Electroencephalography → associated_with → Neurology</td>
<td>0.726124835</td>
</tr>
<tr>
<td>5</td>
<td>Autism → associated_with → Childhood disintegrative disorder</td>
<td>0.706245865</td>
</tr>
<tr>
<td>5</td>
<td>Gyrus → attributed_continuous_with → Postcentral gyrus</td>
<td>0.86015795</td>
</tr>
<tr>
<td>5</td>
<td>Zona_Incerta → attributed_continuous_with → Basal forebrain</td>
<td>0.743588745</td>
</tr>
<tr>
<td>5</td>
<td>Middle_Temporal_Gyrus → attributed_continuous_with → Inferior temporal gyrus</td>
<td>0.70127969</td>
</tr>
<tr>
<td>5</td>
<td>Visual_perception → biological_process → has_associated_location → Brain</td>
<td>0.933201604</td>
</tr>
<tr>
<td>5</td>
<td>Ampa_Receptor → biological_process → has_associated_location → Nervous system</td>
<td>0.8238867</td>
</tr>
<tr>
<td>5</td>
<td>Neurogenesis → biological_process → has_associated_location → Nervous system</td>
<td>0.727214341</td>
</tr>
<tr>
<td>5</td>
<td>Anxiety → definitional_manifestation_of → Anxiety disorder</td>
<td>0.762852273</td>
</tr>
<tr>
<td>5</td>
<td>Demyelinating_Disease → disease → has_associated_anatomic_site → Nervous system</td>
<td>0.855276322</td>
</tr>
<tr>
<td>5</td>
<td>Dopamine_Antagonist → entry_comination_of → Dopamine</td>
<td>0.810267157</td>
</tr>
<tr>
<td>5</td>
<td>Dopaminergic → finding_site_of → Bipolar disorder</td>
<td>0.713380615</td>
</tr>
<tr>
<td>5</td>
<td>Nucleus_Accumbens → gene_plays_role_in_process → Fear</td>
<td>0.737351363</td>
</tr>
<tr>
<td>5</td>
<td>Hypothalamus → has_constitutional_part → Neuroanatomy</td>
<td>0.796539791</td>
</tr>
<tr>
<td>5</td>
<td>Selective_Serotonin_Reuptake_Inhibitor → has_direct_procedure_site → Bipolar disorder</td>
<td>0.738022022</td>
</tr>
<tr>
<td>5</td>
<td>Expressive_Aphasia → has_finding_site → Brain</td>
<td>0.789672998</td>
</tr>
<tr>
<td>5</td>
<td>Alexander_Disease → has_finding_site → Brain</td>
<td>0.708735221</td>
</tr>
<tr>
<td>5</td>
<td>Central_Nervous_System → has_part → Meninges</td>
<td>0.768017609</td>
</tr>
<tr>
<td>5</td>
<td>Nervous_System → has_part → Parasympathetic nervous system</td>
<td>0.749519213</td>
</tr>
<tr>
<td>5</td>
<td>Neuroscience → has_part → Sensory integration</td>
<td>0.72950231</td>
</tr>
<tr>
<td>5</td>
<td>Ganglion → has_physical_part_of_anatomic_structure → Ganglion cell</td>
<td>0.721967102</td>
</tr>
<tr>
<td>5</td>
<td>Motor_Cortex → has_regional_part → Betz cell</td>
<td>0.704433533</td>
</tr>
<tr>
<td>5</td>
<td>Cognitive_Test → has_tradename → Intelligence quotient</td>
<td>0.73906791</td>
</tr>
<tr>
<td>5</td>
<td>Fatal_Familial_Insomnia → interprets → Sleep</td>
<td>0.894434884</td>
</tr>
<tr>
<td>5</td>
<td>Dendritic_Spine → is_associated_anatomic_site_of → Autism</td>
<td>0.789459906</td>
</tr>
<tr>
<td>5</td>
<td>Neural_Development → measures → Autism</td>
<td>0.719980802</td>
</tr>
<tr>
<td>5</td>
<td>Dorsal_Root_Ganglion → part_of → Central nervous system</td>
<td>0.824102775</td>
</tr>
<tr>
<td>5</td>
<td>Spinal_Cord → part_of → Central nervous system</td>
<td>0.788532117</td>
</tr>
<tr>
<td>5</td>
<td>Superior_Temporal_Gyrus → part_of → Cerebral cortex</td>
<td>0.743588094</td>
</tr>
<tr>
<td>5</td>
<td>Visual_Cortex → part_of → Cerebral hemisphere</td>
<td>0.712880668</td>
</tr>
<tr>
<td>5</td>
<td>Human_Brain → process_involves_gene → Bipolar disorder</td>
<td>0.745231561</td>
</tr>
</tbody>
</table>

Continued on next page
### Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Orbitofrontal_Cortex → process_involves_gene → Cognition</td>
<td>0.700973545</td>
</tr>
<tr>
<td>5</td>
<td>Claustrum → receives_input_from → Basal ganglia</td>
<td>0.892184735</td>
</tr>
<tr>
<td>1</td>
<td>Zona_Incerta → receives_input_from → Olfactory bulb</td>
<td>0.950395244</td>
</tr>
<tr>
<td>5</td>
<td>Clinical_Psychology → regional_part_of → Psychoanalysis</td>
<td>0.798546814</td>
</tr>
<tr>
<td>5</td>
<td>Dorsal_Column_Nuclei → regional_part_of → Brain</td>
<td>0.744527049</td>
</tr>
<tr>
<td>5</td>
<td>Visual_Cortex → regional_part_of → Cerebral hemisphere</td>
<td>0.734353006</td>
</tr>
<tr>
<td>5</td>
<td>Paramedian_Pontine_Reticular_Formation → related_to → Brain</td>
<td>0.765186195</td>
</tr>
<tr>
<td>5</td>
<td>Reticular_Activating_System → sib_in_branch_of → Reticular formation</td>
<td>0.872216191</td>
</tr>
<tr>
<td>5</td>
<td>Psychotherapy → sib_in_part_of → Behaviorism</td>
<td>0.871844344</td>
</tr>
<tr>
<td>5</td>
<td>Psycholinguistics → use → Language</td>
<td>0.729197522</td>
</tr>
<tr>
<td>4</td>
<td>Pituitary_Gland → anatomic_structure_has_location → Supraoptic nucleus</td>
<td>0.883394138</td>
</tr>
<tr>
<td>4</td>
<td>Parallel_Processing → anatomic_structure_is_physical_part_of → Brain</td>
<td>0.756607864</td>
</tr>
<tr>
<td>4</td>
<td>Ampa_Receptor → anatomic_structure_is_physical_part_of → Nervous system</td>
<td>0.753646204</td>
</tr>
<tr>
<td>4</td>
<td>Autism → associated_with → Dendritic spine</td>
<td>0.872215083</td>
</tr>
<tr>
<td>4</td>
<td>Psychotherapy → associated_with → Existential therapy</td>
<td>0.769737952</td>
</tr>
<tr>
<td>4</td>
<td>Middle_Temporal_Gyrus → attributed_continuous_with → Superior temporal gyrus</td>
<td>0.903467660</td>
</tr>
<tr>
<td>4</td>
<td>Middle_Frontal_Gyrus → attributed_continuous_with → Superior frontal gyrus</td>
<td>0.885593376</td>
</tr>
<tr>
<td>4</td>
<td>Superior_Frontal_Gyrus → attributed_continuous_with → Inferior frontal gyrus</td>
<td>0.815846561</td>
</tr>
<tr>
<td>4</td>
<td>Gyrus → attributed_continuous_with → Superior frontal gyrus</td>
<td>0.707473131</td>
</tr>
<tr>
<td>4</td>
<td>Inferior_Temporal_Gyrus → attributed_continuous_with → Temporal lobe</td>
<td>0.706509962</td>
</tr>
<tr>
<td>4</td>
<td>Awareness → biological_process_has_associated_location → Brain</td>
<td>0.897154503</td>
</tr>
<tr>
<td>4</td>
<td>Minimum_Alveolar_Concentration → biological_process_involves_gene_product → Brain</td>
<td>0.738628012</td>
</tr>
<tr>
<td>4</td>
<td>Human_Brain → biological_process_involves_gene_product → Bipolar disorder</td>
<td>0.732061396</td>
</tr>
<tr>
<td>4</td>
<td>Working_Memory → biological_process_involves_gene_product → Basal ganglia</td>
<td>0.722428314</td>
</tr>
<tr>
<td>4</td>
<td>Neurodegenerative_Disease → classified_as → Fronto temporal lobar degeneration</td>
<td>0.757826494</td>
</tr>
<tr>
<td>4</td>
<td>Panic_Disorder → classifies → Somatiform disorder</td>
<td>0.841321762</td>
</tr>
<tr>
<td>4</td>
<td>Tourette_Syndrome → clinically_similar → Spectrum disorder</td>
<td>0.783058066</td>
</tr>
<tr>
<td>4</td>
<td>Apathy → clinically_similar → Bipolar disorder</td>
<td>0.746621295</td>
</tr>
<tr>
<td>4</td>
<td>Deep_Brain_Stimulation → gene.plays.role_in.process → Cognition</td>
<td>0.740757749</td>
</tr>
<tr>
<td>4</td>
<td>Mammillary_Body → gene.plays.role_in.process → Papez circuit</td>
<td>0.709121763</td>
</tr>
<tr>
<td>4</td>
<td>Neural_Development → gene_product_has_associated_anatomy → Neural tube</td>
<td>0.719066742</td>
</tr>
<tr>
<td>4</td>
<td>Norepinephrine_Transporter → has_associated_morphology → Neurotransmitter</td>
<td>0.720877273</td>
</tr>
<tr>
<td>4</td>
<td>Action_Learning → has_direct_procedure_site → Learning</td>
<td>0.833427953</td>
</tr>
<tr>
<td>4</td>
<td>Mental_Health_Professional → has_direct_procedure_site → Family therapy</td>
<td>0.792528928</td>
</tr>
<tr>
<td>4</td>
<td>Schizophrenia → has_entry_combination → Mesolimbic pathway</td>
<td>0.751565745</td>
</tr>
<tr>
<td>4</td>
<td>Psychiatry → has_finding_site → Human brain</td>
<td>0.80944437</td>
</tr>
<tr>
<td>4</td>
<td>Schizophreniform_Disorder → has_finding_site → Brain</td>
<td>0.791521035</td>
</tr>
<tr>
<td>4</td>
<td>Eliminative_Materialism → has_finding_site → Brain</td>
<td>0.751777885</td>
</tr>
<tr>
<td>4</td>
<td>Electroencephalography → has_finding_site → Dendrite</td>
<td>0.740699942</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Alexithymia → has_finding_site → Brain</td>
<td>0.720075028</td>
</tr>
<tr>
<td>4</td>
<td>Projection_Fibers → has_laterality → Cerebral cortex</td>
<td>0.706327539</td>
</tr>
<tr>
<td>4</td>
<td>Psychiatry → has_location → Human brain</td>
<td>0.809542661</td>
</tr>
<tr>
<td>4</td>
<td>Cingulate_Gyrus → has_location → Cingulate sulcus</td>
<td>0.73143954</td>
</tr>
<tr>
<td>4</td>
<td>Primary_Motor_Cortex → has_manifestation → Internal capsule</td>
<td>0.738221317</td>
</tr>
<tr>
<td>4</td>
<td>Nervous_System → has_part → Peripheral nervous system</td>
<td>0.8278711</td>
</tr>
<tr>
<td>4</td>
<td>Neuroscience → has_part → Peripheral nervous system</td>
<td>0.827740238</td>
</tr>
<tr>
<td>4</td>
<td>Cisterna_Magna → has_part → Medulla oblongata</td>
<td>0.810250301</td>
</tr>
<tr>
<td>4</td>
<td>Central_Nervous_System → has_part → Spinal cord</td>
<td>0.794698047</td>
</tr>
<tr>
<td>4</td>
<td>Concussion → has_part → Diencephalon</td>
<td>0.763978214</td>
</tr>
<tr>
<td>4</td>
<td>Schizoaffective_Disorder → has_part → Psychotherapy</td>
<td>0.733619319</td>
</tr>
<tr>
<td>4</td>
<td>Intelligence → has_physical_part_of_anatomic_structure → Predictive validity</td>
<td>0.832600338</td>
</tr>
<tr>
<td>4</td>
<td>Medulla_Oblongata → has REGIONAL_part → Medial lemniscus</td>
<td>0.795578674</td>
</tr>
<tr>
<td>4</td>
<td>Limbic_System → has_REGIONAL_part → Septal nuclei</td>
<td>0.75821169</td>
</tr>
<tr>
<td>4</td>
<td>Cognition → ingredient_of → Animal cognition</td>
<td>0.787471092</td>
</tr>
<tr>
<td>4</td>
<td>Doubt → interpretation_of → Logic</td>
<td>0.713395039</td>
</tr>
<tr>
<td>4</td>
<td>Sleep_Deprivation → interprets → Sleep</td>
<td>0.7420927</td>
</tr>
<tr>
<td>4</td>
<td>Mental_Status_Examination → inverse_may_be_a → Anxiety</td>
<td>0.779936181</td>
</tr>
<tr>
<td>4</td>
<td>Language → is_associated_anatomy_of_gene_product → Angular gyrus</td>
<td>0.792257875</td>
</tr>
<tr>
<td>4</td>
<td>Fourth_Ventricle → is_location_of_anatomic_structure → Facial colliculus</td>
<td>0.715114495</td>
</tr>
<tr>
<td>4</td>
<td>Corpus_Striatum → is_location_of_biological_process → Striatum</td>
<td>0.719009959</td>
</tr>
<tr>
<td>4</td>
<td>Selective_Serotonin_Reuptake_Inhibitor → location_of → Dopamine</td>
<td>0.833567396</td>
</tr>
<tr>
<td>4</td>
<td>Primary_Motor_Cortex → manifestation_of → Internal capsule</td>
<td>0.719246147</td>
</tr>
<tr>
<td>4</td>
<td>Primary_Motor_Cortex → may_be_a → Internal capsule</td>
<td>0.728180648</td>
</tr>
<tr>
<td>4</td>
<td>Spinothalamic_Tract → part_of → Spinal cord</td>
<td>0.916179629</td>
</tr>
<tr>
<td>4</td>
<td>Rubrospinal_Tract → part_of → Spinal cord</td>
<td>0.872272085</td>
</tr>
<tr>
<td>4</td>
<td>Dorsal_Column_Nuclei → part_of → Brain</td>
<td>0.860217777</td>
</tr>
<tr>
<td>4</td>
<td>Occipital_Lobe → part_of → Cerebral hemisphere</td>
<td>0.756482789</td>
</tr>
<tr>
<td>4</td>
<td>Autosomal_Dominant_Nocturnal_Frontal_Lobe_Epilepsy → part_of → Frontal lobe</td>
<td>0.723301322</td>
</tr>
<tr>
<td>4</td>
<td>Parallel_Processing → part_of → Brain</td>
<td>0.719937488</td>
</tr>
<tr>
<td>4</td>
<td>Raphe_Nuclei → receives_input_from → Dorsal raphe nucleus</td>
<td>0.816894764</td>
</tr>
<tr>
<td>4</td>
<td>Rubrospinal_Tract → receives_input_from → Red nucleus</td>
<td>0.778827745</td>
</tr>
<tr>
<td>4</td>
<td>Peripheral_Nervous_System → sib_in_part_of → Autonomic nervous system</td>
<td>0.881590041</td>
</tr>
<tr>
<td>4</td>
<td>Incus → sib_in_part_of → Stapes</td>
<td>0.785798286</td>
</tr>
<tr>
<td>4</td>
<td>Malleus → sib_in_part_of → Incus</td>
<td>0.747507163</td>
</tr>
<tr>
<td>4</td>
<td>Central_Nervous_System → sib_in_part_of → Periphera nervous system</td>
<td>0.70077198</td>
</tr>
<tr>
<td>4</td>
<td>Round_Window → use → Middle ear</td>
<td>0.79951174</td>
</tr>
<tr>
<td>4</td>
<td>Schizotypy → use → Schizotypal personality disorder</td>
<td>0.759048264</td>
</tr>
<tr>
<td>4</td>
<td>Cingulate_Gyrus → used_for → Cingulate cortex</td>
<td>0.917447394</td>
</tr>
<tr>
<td>3</td>
<td>Functional_Magnetic_Resonance_Imaging → anatomic_structure_is_physical_part_of → Cerebral cortex</td>
<td>0.71948098</td>
</tr>
<tr>
<td>3</td>
<td>Amnesia → associated_with → Repressed memory</td>
<td>0.707090305</td>
</tr>
<tr>
<td>3</td>
<td>White_Matter → attributed_continuous_with → Fastigial nucleus</td>
<td>0.791355929</td>
</tr>
</tbody>
</table>
Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Insular Cortex → attributed_continuous_with → Parietal lobe</td>
<td>0.772726984</td>
</tr>
<tr>
<td>3</td>
<td>Utilitarianism → biological_process_has_associated_location → Cognitive dissonance</td>
<td>0.727582689</td>
</tr>
<tr>
<td>3</td>
<td>Retrograde_Amnesia → biological_process_involves_gene_product → Declarative memory</td>
<td>0.795205806</td>
</tr>
<tr>
<td>3</td>
<td>Diagnostic And Statistical Manual Of Mental Disorders → classified_as → Epilepsy</td>
<td>0.741808758</td>
</tr>
<tr>
<td>3</td>
<td>Melancholia → classifies → Fear</td>
<td>0.765345216</td>
</tr>
<tr>
<td>3</td>
<td>Intelligence → classifies → Predictive validity</td>
<td>0.758904324</td>
</tr>
<tr>
<td>3</td>
<td>Neuroscience → entry_combination_of → Neural engineering</td>
<td>0.771887086</td>
</tr>
<tr>
<td>3</td>
<td>Selective_Serotonin_Reuptake_Inhibitor → gene_plays_role_in_process → Apathy</td>
<td>0.837197698</td>
</tr>
<tr>
<td>3</td>
<td>Paleopallium → has_component → Brain</td>
<td>0.712525608</td>
</tr>
<tr>
<td>3</td>
<td>Twin_Study → has_direct_procedure_site → Correlation</td>
<td>0.878511591</td>
</tr>
<tr>
<td>3</td>
<td>Subdural_Hematoma → has_direct_procedure_site → Cerebrospinal fluid</td>
<td>0.790542008</td>
</tr>
<tr>
<td>3</td>
<td>Shyness → has_finding_site → Autism</td>
<td>0.885378113</td>
</tr>
<tr>
<td>3</td>
<td>Schizoaffective_Disorder → has_finding_site → Psychotherapy</td>
<td>0.795117407</td>
</tr>
<tr>
<td>3</td>
<td>Neural_Correlate → has_finding_site → Observation</td>
<td>0.780974293</td>
</tr>
<tr>
<td>3</td>
<td>Progressive_Multifocal_Lukenoencephalopathy → has_finding_site → Cerebrospinal fluid</td>
<td>0.748256516</td>
</tr>
<tr>
<td>3</td>
<td>Psychological_Abuse → has_interpretation → Anxiety</td>
<td>0.806023168</td>
</tr>
<tr>
<td>3</td>
<td>Intracranial_Pressure → has_part → Cerebral hemisphere</td>
<td>0.802232689</td>
</tr>
<tr>
<td>3</td>
<td>Limbic_System → has_part → Septal nuclei</td>
<td>0.710365599</td>
</tr>
<tr>
<td>3</td>
<td>Inferior_Colliculus → has_physical_part_of_anatomic_structure → Interaural time difference</td>
<td>0.787360546</td>
</tr>
<tr>
<td>3</td>
<td>Temporal_Lobe → has_regional_part → Parahippocampal gyrus</td>
<td>0.705516981</td>
</tr>
<tr>
<td>3</td>
<td>Schizotypy → interprets → Delusion</td>
<td>0.851619916</td>
</tr>
<tr>
<td>3</td>
<td>Nervous_System → is_associated_anatomic_site_of → Third ventricle</td>
<td>0.847434067</td>
</tr>
<tr>
<td>3</td>
<td>Neuroscience → is_associated_anatomic_site_of → Epilepsy</td>
<td>0.726808414</td>
</tr>
<tr>
<td>3</td>
<td>Creativity → is_interpreted_by → Intelligence quotient</td>
<td>0.832751012</td>
</tr>
<tr>
<td>3</td>
<td>Sleep → is_interpreted_by → Sleep deprivation</td>
<td>0.728782922</td>
</tr>
<tr>
<td>3</td>
<td>Neuroscience → is_location_of_biological_process → Perception</td>
<td>0.90380635</td>
</tr>
<tr>
<td>3</td>
<td>Neuroscience → is_location_of_biological_process → Somatosensory system</td>
<td>0.89480852</td>
</tr>
<tr>
<td>3</td>
<td>Selective_Serotonin_Reuptake_Inhibitor → location_of → Anxiety disorder</td>
<td>0.758144548</td>
</tr>
<tr>
<td>3</td>
<td>Round_Round → part_of → Cochlea</td>
<td>0.771782785</td>
</tr>
<tr>
<td>3</td>
<td>Cerebral_Peduncle → part_of → Corticospinal tract</td>
<td>0.720862711</td>
</tr>
<tr>
<td>3</td>
<td>Subthalamus → receives_input_from → Thalamic reticular nucleus</td>
<td>0.936828649</td>
</tr>
<tr>
<td>3</td>
<td>Claustrum → receives_input_from → Neocortex</td>
<td>0.863380015</td>
</tr>
<tr>
<td>3</td>
<td>Frontal_Lobe → receives_input_from → Internal capsule</td>
<td>0.858006064</td>
</tr>
<tr>
<td>3</td>
<td>Internal_Capsule → receives_input_from → Frontal lobe</td>
<td>0.814602448</td>
</tr>
<tr>
<td>3</td>
<td>Cephalization → regional_part_of → Central nervous system</td>
<td>0.706084792</td>
</tr>
<tr>
<td>3</td>
<td>Oppositional_Defiant_Disorder → related_to → Comorbidity</td>
<td>0.874853018</td>
</tr>
<tr>
<td>3</td>
<td>Caudate_Nucleus → sends_output_to → Globus pallidus</td>
<td>0.803671802</td>
</tr>
<tr>
<td>3</td>
<td>Inner_Ear → sib_in_part_of → Outer ear</td>
<td>0.959002563</td>
</tr>
<tr>
<td>3</td>
<td>Empathy → sib_in_part_of → Proprioception</td>
<td>0.842808881</td>
</tr>
<tr>
<td>3</td>
<td>Subthalamus → sib_in_part_of → Mesencephalon</td>
<td>0.735909066</td>
</tr>
<tr>
<td>3</td>
<td>Mesencephalon → used_for → Red nucleus</td>
<td>0.862911731</td>
</tr>
<tr>
<td>3</td>
<td>Nootropic → used_for → Pyritinol</td>
<td>0.707545429</td>
</tr>
</tbody>
</table>

Continued on next page
Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Score</th>
<th>Triple</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Metencephalon → receives_input_from → Medulla oblongata</td>
<td>0.84898492</td>
</tr>
<tr>
<td>2</td>
<td>Pedunculopontine Nucleus → receives_input_from → Basal ganglia</td>
<td>0.831584387</td>
</tr>
<tr>
<td>2</td>
<td>Subthalamic_Nucleus → receives_input_from → Thalamus</td>
<td>0.829308464</td>
</tr>
<tr>
<td>2</td>
<td>Striatum → receives_input_from → Superior colliculus</td>
<td>0.7863263</td>
</tr>
<tr>
<td>2</td>
<td>Mismatch_Negativity → receives_input_from → Magnetoencephalography</td>
<td>0.784353397</td>
</tr>
<tr>
<td>2</td>
<td>Visual_System → receives_input_from → Cerebral cortex</td>
<td>0.784348028</td>
</tr>
<tr>
<td>2</td>
<td>Subthalamus → receives_input_from → Caudate nucleus</td>
<td>0.761199944</td>
</tr>
<tr>
<td>2</td>
<td>Subthalamus → receives_input_from → Subthalamic nucleus</td>
<td>0.751295453</td>
</tr>
<tr>
<td>2</td>
<td>Subthalamus → receives_input_from → Thalamus</td>
<td>0.723864812</td>
</tr>
<tr>
<td>2</td>
<td>Sensory_System → regional_part_of → Somatosensory system</td>
<td>0.831628224</td>
</tr>
<tr>
<td>2</td>
<td>Entorhinal_Cortex → regional_part_of → Parahippocampal gyrus</td>
<td>0.744738683</td>
</tr>
<tr>
<td>2</td>
<td>Neuron → related_to → Medium spiny neuron</td>
<td>0.907110249</td>
</tr>
<tr>
<td>2</td>
<td>Limbic_System → sends_output_to → Pituitary gland</td>
<td>0.866987095</td>
</tr>
<tr>
<td>2</td>
<td>Frontal_Lobe → sends_output_to → Internal capsule</td>
<td>0.836698843</td>
</tr>
<tr>
<td>2</td>
<td>Internal_Capsule → sends_output_to → Frontal lobe</td>
<td>0.836438143</td>
</tr>
<tr>
<td>2</td>
<td>Amygdalofugal_Pathway → sends_output_to → Subthalamic nucleus</td>
<td>0.807346604</td>
</tr>
<tr>
<td>2</td>
<td>Inferior_Colliculus → sends_output_to → Lateral geniculate nucleus</td>
<td>0.770382416</td>
</tr>
<tr>
<td>2</td>
<td>Superior_Colliculus → sends_output_to → Spinal trigeminal nucleus</td>
<td>0.730063901</td>
</tr>
<tr>
<td>2</td>
<td>Septal_Nuclei → sends_output_to → Amygdala</td>
<td>0.707499847</td>
</tr>
<tr>
<td>2</td>
<td>Postcentral_Gyrus → sib_in_part_of → Insular cortex</td>
<td>0.839184109</td>
</tr>
<tr>
<td>2</td>
<td>Problem_Solving → subtype_of → Psychologist</td>
<td>0.917417511</td>
</tr>
<tr>
<td>2</td>
<td>Sympathy → subtype_of → Grief</td>
<td>0.866895436</td>
</tr>
<tr>
<td>1</td>
<td>Depersonalization → classifies → Depersonalization disorder</td>
<td>0.924717497</td>
</tr>
<tr>
<td>1</td>
<td>Selective_Serotonin_Reuptake_Inhibitor → classifies → Mania</td>
<td>0.866484752</td>
</tr>
<tr>
<td>1</td>
<td>Interpeduncular_Cistern → disease_has_associated_anatomic_site → Cerebral peduncle</td>
<td>0.887699884</td>
</tr>
<tr>
<td>1</td>
<td>Trichotillomania → disease_has_anatomic_site → Kleptomania</td>
<td>0.716794259</td>
</tr>
<tr>
<td>1</td>
<td>Paramedian_Pontine_Reticular_Formation → gene_product_has_biochemical_function → Brain</td>
<td>0.911483563</td>
</tr>
<tr>
<td>1</td>
<td>Cognitive_Neuropsychiatry → gene_product_has_biochemical_function → Psychopathology</td>
<td>0.838076941</td>
</tr>
<tr>
<td>1</td>
<td>Progressive_Multifocal_Leukoencephalopathy → gene_product_plays_role_in_biological_process</td>
<td>0.854302341</td>
</tr>
<tr>
<td></td>
<td>→ White matter</td>
<td>0.853403241</td>
</tr>
<tr>
<td>1</td>
<td>Deep_Brain_Stimulation → gene_product_plays_role_in_biological_process → Cognition</td>
<td>0.837485848</td>
</tr>
<tr>
<td>1</td>
<td>Classical_Conditioning → is_primary_anatomic_site_of_disease → Interstimulus interval</td>
<td>0.760498993</td>
</tr>
<tr>
<td>1</td>
<td>Repressed_Memory → is_primary_anatomic_site_of_disease → Amnesia</td>
<td>0.760405599</td>
</tr>
<tr>
<td>1</td>
<td>Conduct_Disorder → may_be_a → Conduct disorder</td>
<td>0.760979762</td>
</tr>
<tr>
<td>1</td>
<td>Superior_Colliculus → part_of → Vertebrate</td>
<td>0.895388444</td>
</tr>
<tr>
<td>1</td>
<td>Lateral_Reticular_Formation → receives_input_from → Thalamus</td>
<td>0.956443571</td>
</tr>
<tr>
<td>1</td>
<td>Globus_Pallidus → receives_input_from → Subthalamic nucleus</td>
<td>0.886636864</td>
</tr>
<tr>
<td>1</td>
<td>Subthalamus → receives_input_from → Putamen</td>
<td>0.882665886</td>
</tr>
<tr>
<td>1</td>
<td>Primary_Auditory_Cortex → regional_part_of → Schizophrenia</td>
<td>0.715758526</td>
</tr>
<tr>
<td>1</td>
<td>Zona_Incerta → sends_output_to → Cingulate cortex</td>
<td>0.857502093</td>
</tr>
</tbody>
</table>