From Text to Knowledge: Bridging the Gap with Probabilistic Graphical Models

Markus Bundschus

München 2010

From Text to Knowledge: Bridging the Gap with Probabilistic Graphical Models

Markus Bundschus

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Markus Bundschus
aus München

München, den 01.06.2010

Erstgutachter: Prof. Dr. Hans-Peter Kriegel
Zweitgutachter: Prof. Dr. Philipp Cimiano
Drittgutachter: Dr. Volker Tresp
Tag der mündlichen Prüfung: 21.07.2010

Contents

Acknowledgments

During the last three years, many individuals advised, guided and supported me in many different ways and thus greatly contributed to this thesis. I am deeply grateful for this.

First of all, I would like to thank Prof. Hans-Peter Kriegel. Without him, this thesis would not have been possible and he consistently supported me during my research. It was a great honor for me to be a member of such a successful research group.

I am also greatly thankful to Prof. Philipp Cimiano who kindly agreed to examine this thesis and spends his extremely spare time for this. Also many thanks to Prof. François Bry and Prof. Martin Wirsing for being part of my dissertation committee.

I owe my deepest thanks to Dr. Volker Tresp, who guided and advised me during the last three years. Volker always encouraged me with his positive attitude and inspiring words. I was very fortunate to have the opportunity to work with him. His infectious enthusiasm for research, open-mindedness and perceptive thoughts have been a major factor in my research.

I am very grateful to my friends and former colleagues Dr. Mathäus Dejori and Dr. Shipeng Yu. Both contributed significantly to my research. Special thanks to Shipeng, who gave me the opportunity to work in the CAD & Knowledge Solutions Group at Siemens Healthcare in the US.

There are, of course, many other people whose advice, discussions, comments and feedback have greatly contributed to my Ph.D. work, both directly and indirectly. I would like to thank Dr. Florian Steinke, Dr. Thomas Runkler, Prof. Martin Greiner, Maximilian Nickel, Thorsten Führing, Stefan Weber, Christa Singer, Davide Magatti, Yi Huang, Peer Kröger, Matthias Schubert, Susanne Grienberger and Franz Krojer.

Finally, I am deeply grateful to my family and friends for their never ending support and their love. My deepest thanks to Nicoline for all her support and care during the last months.

Abstract

The global information space provided by the World Wide Web has changed dramatically the way knowledge is shared all over the world. To make this unbelievable huge information space accessible, search engines index the uploaded contents and provide efficient algorithmic machinery for ranking the importance of documents with respect to an input query. All major search engines such as Google, Yahoo or Bing are keyword-based, which is indisputable a very powerful tool for accessing information needs centered around documents. However, this unstructured, document-oriented paradigm of the World Wide Web has serious drawbacks, when searching for specific knowledge about real-world entities. When asking for advanced facts about entities, today's search engines are not very good in providing accurate answers. Hand-built knowledge bases such as Wikipedia or its structured counterpart DBpedia are excellent sources that provide common facts. However, these knowledge bases are far from being complete and most of the knowledge lies still buried in unstructured documents.

Statistical machine learning methods have the great potential to help to bridge the gap between text and knowledge by (semi-)automatically transforming the unstructured representation of the today's World Wide Web to a more structured representation. This thesis is devoted to reduce this gap with Probabilistic Graphical Models. Probabilistic Graphical Models play a crucial role in modern pattern recognition as they merge two important fields of applied mathematics: Graph Theory and Probability Theory.

The first part of the thesis will present a novel system called *Text2SemRel* that is able to (semi-)automatically construct knowledge bases from textual document collections. The resulting knowledge base consists of facts centered around entities and their relations. Essential part of the system is a novel algorithm for extracting relations between entity mentions that is based on *Conditional Random Fields*, which are Undirected Probabilistic Graphical Models.

In the second part of the thesis, we will use the power of Directed Probabilistic Graphical Models to solve important knowledge discovery tasks in semantically annotated large document collections. In particular, we present extensions of the *Latent Dirichlet Allocation* framework that are able to learn in an unsupervised way the statistical semantic dependencies between unstructured representations such as documents and their semantic annotations. Semantic annotations of documents might refer to concepts originating from a thesaurus or ontology but also to user-generated informal tags in social tagging systems. These forms of annotations represent a first step towards the conversion to a more structured form of the World Wide Web.

In the last part of the thesis, we prove the large-scale applicability of the proposed fact extraction system *Text2SemRel*. In particular, we extract semantic relations between genes and diseases from a large biomedical textual repository. The resulting knowledge base contains far more potential disease genes exceeding the number of disease genes that are currently stored in curated databases. Thus, the proposed system is able to unlock knowledge currently buried in the literature. The literature-derived human gene-disease network is subject of further analysis with respect to existing curated state of the art databases. We analyze the derived knowledge base quantitatively by comparing it with several curated databases with regard to size of the databases and properties of known disease genes among other things. Our experimental analysis shows that the facts extracted from the literature are of high quality.

Zusammenfassung

Das Internet hat sich zu einem Informationsraum entwickelt, der die Art und Weise wie Wissen global zur Verfügung gestellt wird, dramatisch verändert hat. Um diesen umfangreichen Informationsraum zugänglich zu machen, indexieren Suchmaschinen global verteilte Inhalte. Zudem bieten sie effiziente algorithmische Werkzeuge an, Dokumente bezüglich einer Abfrage der Relevanz nach zu sortieren. Alle bekannten Suchmaschinen wie beispielsweise Google, Yahoo oder Bing ermöglichen eine Suche auf Basis von Schlüsselwörtern. Ein nicht mehr wegzudenkendes und mächtiges Werkzeug — allerdings nur, wenn das Ziel einer Suche die Dokumente selbst sind. Das Dokumenten-zentrierte Paradigma des Internets hat jedoch erhebliche Schwachstellen, wenn es darum geht, spezifisches Wissen über Entitäten der realen Welt zu recherchieren. Sobald man sich auf die Suche nach nicht trivialen Fakten über Entitäten begibt, liefern Suchmaschinen nach momentanem Stand der Technik keine zufriedenstellenden Ergebnisse. Von Menschenhand geschaffene Wissensbasen wie beispielsweise Wikipedia oder das strukturierte Pendant DBpedia sind zwar exzellente Quellen für allgemeine Fakten, jedoch sind sie weit davon entfernt vollständig zu sein. Ein Großteil des Wissens liegt also weiterhin in unstrukturierter Form brach.

Statistisches Maschinelles Lernen kann einen wesentlichen Beitrag leisten das Internet in seiner unstrukturierten Form, (semi-)automatisch in eine strukturierte Form zu überführen und somit die Lücke zwischen unstrukturierter und strukturierter Repräsentation zu schließen. Die vorliegende Dissertation versucht, diese Lücke mit Hilfe Probabilistischer Graphischer Modelle zu reduzieren. Diese Modelle spielen eine zentrale Rolle in der modernen Mustererkennung, da sie zwei wichtige Gebiete der angewandten Mathematik vereinen: die Graphentheorie und die Wahrscheinlichkeitslehre.

Im ersten Teil dieser Arbeit präsentieren wir ein neuartiges System, *Text2SemRel*, das (semi-)automatisch strukturierte Wissensbasen aus Textkollektionen erstellen kann und folglich den Weg für strukturiertes Wissen ebnet. Durch *Text2SemRel* entstandene Wissensbasen bestehen aus Fakten über Entitäten und ihren Relationen untereinander. Grundlegender Bestandteil des Systems ist ein innovativer Algorithmus der Relationen zwischen Entitäten in Text erkennt. Der Algorithmus basiert auf *Conditional Random Fields*, einer speziellen Form sogenannter ungerichteter Probabilistischer Graphischer Modelle.

Im zweiten Teil der Arbeit bedienen wir uns der Mächtigkeit gerichteter Probabilistischer Graphischer Modelle, um aus semantisch annotierten, großen Textkollektionen vorher nicht bekanntes Wissen zu erschließen. Im Besonderen werden Modellerweiterungen präsentiert, die auf der Technik der *Latent Dirichlet Allocation* basieren. Diese Erweiterungen sind in der Lage, statistische semantische Abhängigkeiten zwischen unstrukturierten Formen, beispielsweise Dokumenten, und ihren semantischen Annotationen unüberwacht zu lernen. Mit semantischen Annotationen sind Konzepte aus einem Thesaurus oder aus einer Ontologie zu verstehen. Die Annotationen können aber auch aus sozialen Taggingsystemen in Form von informellen Tags stammen. Semantisch annotierte Dokumente stellen einen ersten wichtigen Schritt in Richtung einer strukturierten Form des Internets dar.

Der letzte Teil der Arbeit behandelt die Anwendbarkeit von *Text2SemRel* auf großen Textsammlungen. Im Speziellen verwenden wir *Text2SemRel*, um semantische Relationen zwischen Genen und Krankheiten aus einer kompletten biomedizinischen Textkollektion zu extrahieren. Die so entstandene Wissensbasis enthält weit mehr Gene, die mit menschlichen Krankheiten in Verbindung gebracht werden, als momentan in manuell kurierten Datenbanken zu finden ist. Damit zeigen wir, dass *Text2SemRel* in der Lage ist, vorher nicht in strukturierter Form verfügbares Wissen aus großen Textdatenbanken zu extrahieren. Das aus der Literatur gewonnene humane Gen-Krankheits-Netzwerk wird mehreren kurierten, modernen Datenbanken gegenübergestellt. Unter anderem vergleichen wir das aus der Literatur gewonnene humane Gen-Krankheits-Netzwerk quantitativ gegen die manuell kurierten Datenbanken in Hinblick auf die Größe der Datenbanken und in Hinblick auf die Charakteristika von bekannten Krankheitsgenen. Anhand der experimentellen Auswertungen können wir zeigen, dass die extrahierten Fakten von hoher Qualität sind.

Chapter 1

Introduction

1.1 Motivation

'Knowledge is power.'
SIR FRANCIS BACON — 1597

Our society is emerging towards a knowledge society in which wealth will be produced mainly by knowledge and not by material goods anymore [72]. An essential asset of the knowledge society represent knowledge workers, individuals whose main task is to interpret complex information. One of the main challenges in the 21st century will be the task to manage knowledge worker productivity [73]. While some knowledge resides within individuals in form of know how or expertise, a form of knowledge that cannot be easily transferred, a huge fraction refers to knowledge that can be easily written down, e. g. in form of manuals, guides or scientific papers. Beyond doubt, the most significant instrument to transfer or communicate knowledge is writing. Therefore, knowledge workers spend most of the time on reading and interpreting written (digital) information.

With the advent of the World Wide Web (WWW), not only knowledge workers, information seekers in general, have to cope with an unprecedented wealth of information. To make this unbelievable huge information space accessible, search engines index the uploaded contents and provide efficient algorithmic machinery for ranking the importance of documents with respect to an input query. In a multitude of cases, however, information needs are centered around facts about real-world entities, an information need which is not covered adequately by today's search engines. Today's search engines are geared towards the document-centered paradigm of the WWW and thus are designed to return documents given an input query. Retrieving complete lists of facts is only one apparent example, where classical search engines fail. In the majority of cases, the information seeker has to browse through many web pages to gather the desired list of facts. If one is lucky and among the highest ranked documents is a web page that lists the desired facts, there is no evidence about the completeness of the list. Indeed, common facts can be found in knowledge bases such as Wikipedia or its structured counterpart, DBpedia, but (i) these are far from being complete and (ii) the desired facts will be often highly specific knowledge that is not stored in these resources. As a consequence, information seekers will have to spend substantial time in order to search in unstructured sources for the desired facts.

This thesis is concerned with knowledge that is communicated in written form and we will investigate ways how knowledge can be (semi-)automatically extracted from unstructured text in order to help to reduce the gap between text and explicit knowledge.

Towards a Paradigm Shift in the Retrieval Unit: From Documents to Entities and Relationships

All major search engines, i. e. Google, Microsoft's Bing and Yahoo, provide evidence that a change from the document-centered paradigm of the WWW towards a more concept-based or entity-based paradigm might be forthcoming in the near future:

- **Yahoo:** In recent position papers [65, 13], the next possible transformative step in the evolution of the web is presented, the web of concepts. The central task in this web will be to rank information about concepts based on the user's information need. The search engine provider motivates the need of a more structured or more semantic web by analyzing query logs [65]. Based on the analysis over a time period of one month, the authors estimate that in approximately 60%-70% of queries, users are looking for specific instances of concepts.

- **Microsoft:** Similar to the intention above, this search engine provider defines web objects as the basic retrieval unit [144]. The corresponding research prototype EntityCube[1] collects and summarizes information about entities. As another concrete example, the Microsoft Academics Search service[2] is based on these technologies.

- **Google:** Also the leading search engine provider has a research prototype that follows up the trend of a concept-based web. Google Squared[3] can be best described as a general web-scale entity and relation extraction tool, that returns given an input query, the most likely entities with corresponding relationships.

Besides this evidence from leading search engine providers, the Linked Data (LD) [23] initiative, also referred to as Web of Data, launched by the father of the WWW itself, Sir Tim Berners-Lee, is the strongest indicator that the nature of the web is undergoing a radical change. Linked Data refers to a set of principles, how to publish data on the web. The Linking Open Data[4] project keeps track of data sets that follow the Linked Data principles. Figure 1.1[5] gives an overview of data sets that already follow these principles (as of June 2009). Nodes in the figure refer to distinct data sets and edges indicate direct links between data sets. In [23], the size of the cloud, measured in number of Resource

[1] http://research.microsoft.com/en-us/projects/entitycube/

[2] http://academic.research.microsoft.com/

[3] http://www.google.com/squared

[4] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[5] Figure taken from http://richard.cyganiak.de/2007/10/lod/

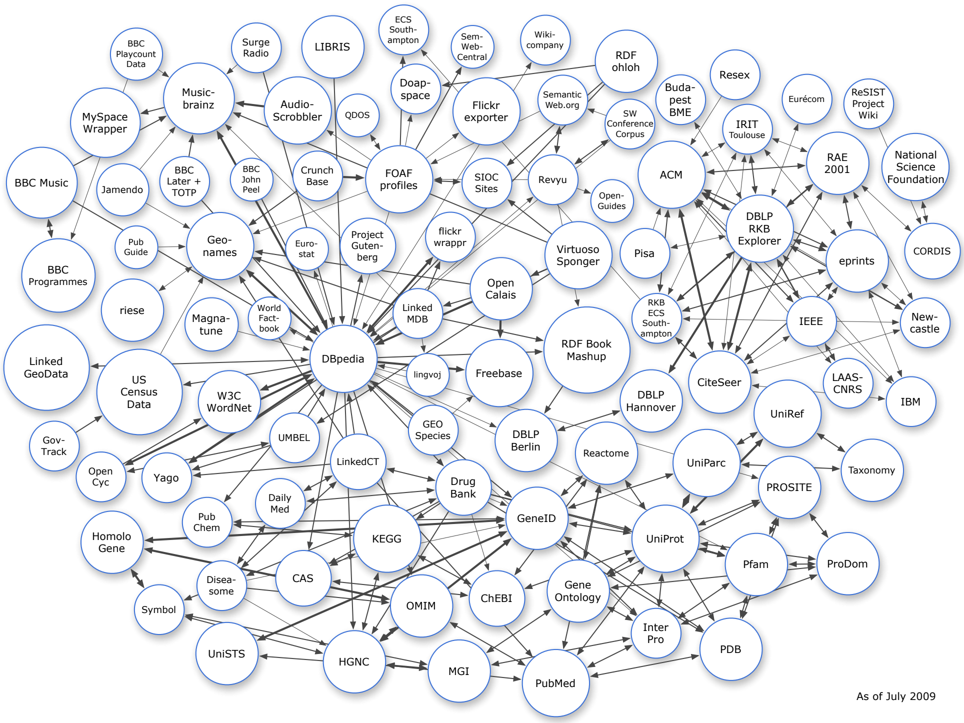


Figure 1.1: **Linked Open Data cloud.**

Description Framework[6] (RDF) triples, was estimated to be about 4.7 billion triples. As of Mai 20th 2010, the number of triples is estimated to be about 13 billion triples[7]. This huge increase clearly indicates that the web of data is active and growing rapidly.

Opportunities and Challenges

Challenges Most of the knowledge still lies buried in unstructured textual sources. We outline the main problems that have to be solved in order to reduce the gap between knowledge encoded in unstructured text and knowledge in explicit form. We do not explicitly discuss challenges that arise after knowledge has been extracted such as trust, redundancy, and noise. By taking a look at the previous section, it quickly becomes clear that entities and relationships take the center stage. All outlined projects, though using slightly different names, choose real-world objects as the basic retrieval unit. The main problems to solve are in particular:

- **Entity extraction:** Identifying the phrases in the text, that refer to real-world objects is the first important task to tackle. This usually comprises the classification of the phrase with respect to the type of entity.

[6] http://www.w3.org/RDF/

[7] http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

- **Entity resolution:** Entities are polysemous, i. e. different textual phrases refer to the same entity. To keep knowledge bases consistent, it is important to solve the resolution of entities.

- **Relationship extraction:** To spin a web of data, we need to identify links between real-world objects. In the majority of cases, relationships between entities are stated in unstructured sources. To improve preciseness, it is important to characterize the links semantically according to the type of relation. Simple co-occurrence of real-world objects inside textual phrases does usually not suffice, because real-world objects can co-occur for many reasons.

- **Classification of unstructured sources:** Besides the just outlined techniques for harvesting knowledge, meta data, originating from controlled vocabularies, which describe the content of documents concisely, will contribute to a more structured form of the Web. Very often the controlled vocabularies will consist of thousands of instances, making accurate text classification quite challenging.

All in all, challenging tasks have to be solved and there are major efforts ongoing in several research communities that try to solve these pressing problems. Nevertheless, it is worth to tackle the challenges, since the opportunities resulting from a more structured representation of unstructured data sources are numerous and very promising.

Opportunities

- **Traditional information retrieval:** Assume that web documents are annotated with semantic meta data, such as facts, entities or concepts. This would significantly offer new filtering opportunities for classical search engines and enhance precision dramatically. As a concrete example, let's assume a knowledge worker is interested in articles discussing a specific protein-protein interaction. With rich semantically annotated documents, one can simply filter for articles discussing the desired interaction.

  But also the presentation of search results of traditional search engines can be improved by augmenting the results with meta data [65]. E. g. , when the result is a page about a restaurant, additional important information such as the location, ratings and opening times can be automatically displayed. The burden for the user of searching the desired information directly on the web page is omitted.

- **Entity-oriented search**: Entities and facts about them are often mentioned countless times in different resources. Entity-oriented search summarizes all this information per entity or per fact. This has the huge advantage that users do not have to search over possibly large amounts of relevant documents in order to assemble the desired information, instead the answer comes in aggregated form. The entities can be represented as a graph, where nodes refer to the entities themselves and edges state relationships between them. With this graph-based view, powerful querying capabilities are enabled over entities and their relationships (see e. g. [113]).

Considered from an application perspective, the entity-oriented view will be also beneficial in a company environment, e. g. for expertise search. Moreover, most of the knowledge in a company exists in unstructured form, therefore semantically annotated documents have direct implications for managing a company's knowledge [184].

- **Data management:** If a crucial mass of organizations publish data according to the LD principles by using existing Uniform Resource Identifiers (URIs) from other data publishers, the creation of further data silos is prevented. Especially in the biomedical domain, where lots of different identifiers exist for the same entities, data silos and inconsistencies are a daily occurrence.

  Once again, this has direct implications for companies. Companies have to make decisions based on internal as well as external data. If a critical mass of knowledge is available in a web of data, companies can connect to this data by using the linked data approach. In this way, additional knowledge can be easily assembled to support the decision making process.

- **Knowledge discovery:** A more structured representation of the Web has direct implications for knowledge discovery. The graph-based representation of entities and their relationships allows to apply powerful graph mining algorithms. Besides this, the network structure of entities can be used to draw conclusions about the network as a whole. For instance, we can easily become a better understanding about the importance of entities and relationships in the network.

1.2 Probabilistic Graphical Models

Statistical machine learning methods have the great potential to help to bridge the gap between text and knowledge by (semi-)automatically transforming unstructured representations to a more structured representation. This thesis is devoted to reduce this gap with Probabilistic Graphical Models (PGMs), which play a crucial role in modern pattern recognition as they merge two important fields of applied mathematics: Graph Theory and Probability Theory. PGMs use a graph-based representation to summarize complex probability distributions in a compact way. In a PGM, the nodes correspond to random variables and the edges encode probabilistic dependencies. PGMs convert the knowledge and assumptions we make about a real-world process into a formal mathematical representation. Hereby, PGMs can deal with uncertainty, a frequently occurring property in real-world applications, since most of the time, we observe only partial and noisy observations. There are two major classes of PGMs: (i) Directed Graphical Models or Bayesian Networks and (ii) Undirected Graphical Models or Markov Random Fields (MRFs). Both classes of models will be employed in this thesis. In the first class of models, the graph is directed, i. e. we have directed links between random variables. This model class allows us to pose interesting probabilistic queries, which can be used for knowledge discovery purposes (see Chapter 3). In contrast, the graphs in MRFs are undirected. A special form of MRFs are Conditional Random Fields which are more appropriate for producing discriminative classifiers (see Chapter 2). A detailed discussion about PGMs in general can be found in [22].

1.3 Contributions of the Thesis

This thesis contributes to advanced forms of extracting information from textual sources. Our main contributions are the following:

- *Text2SemRel* [34].
  We present a new system that (semi-)automatically extracts entities and semantic relations from unstructured data sources. The system is based on Conditional Random Fields, a special type of Undirected Graphical Models. The resulting facts are stored in a knowledge base. *Text2SemRel* comes with an easy to use, graph-based visualization framework, which supports interactive exploration with simple keyword search.

- Topic Models for Semantically Annotated Document Collections [35, 36, 37, 138].
  Web documents are increasingly annotated with semantic meta data. We develop several models that are suitable to model statistical dependencies between the unstructured document representations and their structured annotations. The developed models are based on a special type of Bayesian network and represent a framework in which a number of important knowledge discovery tasks in semantically annotated document collections can be solved.

- LHGDN (**L**iterature-derived **H**uman **G**ene-**D**isease **N**etwork) [83].
  We apply *Text2SemRel* to a challenging biomedical use case in order to prove the power and large-scale applicability of the proposed system. The resulting knowledge base, the LHGDN, is compared with several state of the art gene-disease association repositories. The LHGDN provides high-quality facts about genes and diseases and is, to the best of our knowledge, the largest gene-disease association repository publicly available. The LHGDN is now integral part of the Linked Life Data[8] (LLD) initiative.

Outline The rest of the thesis is structured as follows: Chapter 2 introduces our new information extraction system *Text2SemRel*. Chapter 3 presents probabilistic topic models for knowledge discovery in semantically annotated document collections. Our last contribution of the thesis, the construction of the LHGDN, is presented in Chapter 4. Each of these chapters comes with a motivation and detailed discussion of related work. Moreover,

[8] http://linkedlifedata.com/sources

each chapter ends with a conclusion. Finally, we summarize the thesis and give a brief outlook in Chapter 5.

Chapter 2

Automatic Construction of Knowledge Bases from Textual Data

2.1 Overview

Our objective is to set up a general framework, which automatically constructs a knowledge base from a text collection. Hereby, the knowledge base stores information in form of a so-called Entity-Relationship graph (ER graph), which consists of facts about entities using typed relations (e. g. *[Angela Merkel, born_in, Hamburg]*). Section 2.1.1 gives a detailed description about the used terminology. The extracted data should naturally be of the highest quality possible. As discussed in Section 1.1, the conversion of facts from unstructured document collections into a structured representation, such as an ER graph, offers new, exciting opportunities for search and retrieval.

In this chapter, we present *Text2SemRel* [34], a new approach to build automatically a knowledge base consisting of entities and relations extracted from textual data. Facts can be converted into a canonical form provided that controlled vocabularies of entities are available. By canonical form we mean that the extracted facts are represented by clearly defined relations and unique identifiers for entities [176]. From a semantic web perspective, *Text2SemRel* can be used to populate an ontology of interest. This is a very important prerequisite to make the assembled knowledge available for semantic-based search engines. After *Text2SemRel* has successfully converted a loose text collection, a semantic search engine can exploit the assembled knowledge base. The knowledge base can either be used to enrich an already existing knowledge base or can be employed alone. Every graph-based query language of choice (e. g. the query language used by NAGA, **N**ot **A**nother **G**oogle **A**nswer [113]) can then be used to pose powerful queries to retrieve advanced semantic information. Our method builds upon a new approach for Semantic Relation Extraction (SRE), which is based on Conditional Random Fields (see Section 2.2). We present quantitative results for two biomedical use cases against human generated gold standards:

1. The extraction of typed gene-disease relations from GeneRIF (Gene Reference Into

Function [135]) phrases.

2. The extraction of typed disease-treatment relations from PubMed abstracts.

More important, we demonstrate the full power of *Text2SemRel* by constructing a knowledge base with typed gene-disease relations for the human organism. A biologically motivated discussion about the properties and the quality of this network is presented in Chapter 4. A graph-based, innovative visualization framework is presented, which enables the fast and interactive exploration of the knowledge base.

2.1.1 Terminology

We consider an *entity* to be an instance of a certain *entity class* such as people, companies, diseases, genes and proteins. E. g. , *Angela Merkel* is an instance of the entity class *person*. Two entities can stand in a *relation*. A relation is restricted to hold between entities of two entity classes, i. e. we only consider binary relations. To state that Angela Merkel was born in Hamburg, we say that *Angela Merkel* stands in the *born_in* relation with *Hamburg*, which is written as: *[Angela Merkel, born_in, Hamburg]*.
A relation is well-defined by its *label*, *domain* and *range*. While the label represents simply the name (such as *born_in*), the domain and range narrow down the scope of application of the relation. E. g. the relation *born_in* holds only between the entity classes person and location.

Following the convention in [176], a *triple* consisting of an entity, a relation and another entity is referred to as *fact*. We say that facts are in *canonical form* if all relations are uniquely defined and exactly one identifier is used for an entity [176]. As we will see later, *Text2SemRel* builds upon a *semantic data model* or *conceptual schema*, which we understand as a description about entities and their relationships in the real world. This description might come in form of an Entity-Relationship model [51] or another form of formal description such as an ontology. We distinguish between the type-level or concept-level representation of an ontology and the instance-level. While the latter representation models named entities (i. e. individuals) and their relationships, the type-level aims at modeling classes of entities and their relationships.

A *knowledge base* consists of facts centered around entities and relations and can be represented as an *Entity-Relationship graph* (ER graph), a graph that connects different entities E (E, a finite set of entity labels) through different relations R (R, a finite set of relation labels). Thus, nodes correspond to objects and are labeled with the URI of the object. Directed links are relations and are labeled with the URI of the relation class.

2.1.2 Motivation and Problem Statement

Motivation

The World Wide Web (WWW) has emerged to form the largest information repository of the world, but in its current form it is document-centered. All major search engines such as *Google*, *Yahoo* or *Bing* are still keyword-based, which is indisputable a very powerful tool for accessing information needs centered around documents. But already in this classical information retrieval context, it would be a great benefit to annotate documents with facts about entities and relations between them. This would significantly offer new filtering opportunities for classical search engines and enhance precision dramatically. As a concrete example, let's assume a researcher is interested in articles discussing a specific protein-protein interaction. With rich semantically annotated documents, the researcher can simply filter for articles discussing the desired interaction. Note that the mere co-occurrence of entities is often not precise enough for getting the desired information. For instance, proteins do co-occur in the same document for many reasons, but most of the time a biomedical researcher is interested in articles which state that there is a specific physical interaction between particular proteins.

Furthermore, the unstructured, document-oriented paradigm of the WWW has severe drawbacks, when searching for specific knowledge about real-world entities (see Section 1.1). When asking for advanced facts about entities and relationships, today's search engines are not very good in providing answers for such information needs [113]. Gathering complete lists of facts is another apparent task, where today's search engines fail because of the document-oriented paradigm. There exists already a lot of information about real-world entities on the web in structured or almost structured form, but first, this information is far from being complete, second, most of the time, this information covers only quite common facts and third the desired information might be distributed across several heterogeneous sources. Knowledge repositories such as *Wikipedia* store a lot of facts in almost structured data such as HTML tables as Wikipedia infoboxes, but these tables mainly cover only information about the most common entities and relations (e. g. facts about companies, people and locations). More uncommon, specific entities and relations are often not less important but are still locked in textual form. For instance, while common relations such as *birth_date* between a person entity and a date are often found in the Wikipedia infoboxes, more specific relations (e. g. *graduated_in_year*) are only available in the unstructured Wiki page. Moreover, the relative importance of entities and relations is highly subjective to the end-user's interest. Since not all different information needs can be considered when collecting facts about entities and relations, methods are needed which can extract the desired facts directly at the textual level.

Prime Example: The Biomedical Domain Here, the last decade has seen an explosion of literature. The main reason is the appearance of new biomedical research tools and methods such as high-throughput experiments based on DNA microarrays. The increasing amount of published literature in biomedicine represents an immense source of knowledge. In recent years, it quickly became clear that this overwhelming amount of biomedical literature could only be managed efficiently with the help of automated text information extraction methods. [62] introduce the notion of *knowledge pockets*, meaning that the visibility of facts known to individual researchers appears to be very restricted compared to the whole accessible knowledge. [62] estimate that approximately at least a billion non-redundant molecular interactions are currently locked in the biomedical literature.

As a concrete example, even though there are a lot of data repositories available on the web, which store facts about gene-disease relations, getting a list of all genes which are jointly involved in the development of two or more diseases is extremely tedious. The repositories often have a noticeable delay in updating and are far from being complete [146]. If we are looking for specific typed relations (e. g. *altered_expression*) between genes and diseases, the situation is getting even more complicated. In the worst case, no structured repository has considered this semantic relation so far. The importance of gathering facts from the literature is self-evident here.

Problem Statement

Given typed entities and relations, we would like to be able to infer from an unstructured text collection, the main problem considered here is how to extract entities and relations from a loose text collection and how to arrange the extracted facts in an ER graph.

This problem comes with the following main challenges:

1. Extraction of entities from unstructured text.
2. Extraction of typed relations between entities from unstructured text.
3. Normalization of entities and relations to Uniform Resource Identifiers (URIs).

2.1.3 Proposed Approach at a Glance

Here, we briefly summarize the proposed framework. Full details are given in Section 2.2. As can be seen from Figure 2.1, *Text2SemRel* consists of four main layers:

- **Pre-Processing Layer:** Our method builds on supervised learning, i. e. *Text2SemRel* depends on labeled training data. A corpus of documents has to be annotated with typed entities and relations. Without going into details at this point, entity classes and relations between entity classes originating from a given conceptual schema or semantic data model are aligned with a document collection of interest in the pre-processing phase. We argue that *Text2SemRel* is quite general and can easily be extended to a new domain provided that labeled training data is available. Indeed, the existence of labeled data is often seen as a bottleneck in learning from textual sources and may hamper the adaption of the technique to a new domain. However, with the advancement of professional text annotation services such as *ForScience*[161] or *Amazon Mechanical Turk*[1], this bottleneck might be more and more vanishing in the near future. The work of [161, 48] are recent successful examples of how to outsource this tedious process.

[1] https://www.mturk.com/

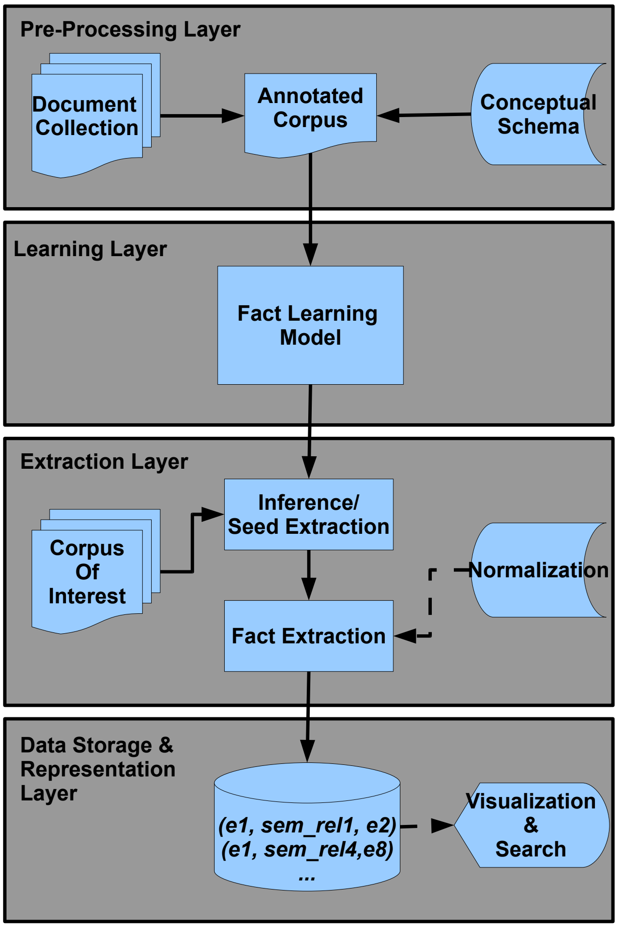


Figure 2.1: **Text2SemRel framework.** The system consists of four main layers: the pre-processing layer, the learning layer, the extraction layer and the data-storage & representation layer. Note that the dotted lines indicate optional steps.

- **Learning Layer:** This represents the most important part of the system. Our approach is based on Conditional Random Fields (CRFs), which are a special type of Probabilistic Graphical Models (PGMs). The fact learning module comprises two challenges: Learning feature weights for extracting named entities from text (Named Entity Recognition (NER)) and learning feature weights for extracting typed relations between them (referred to as Relation Extraction (RE) or Semantic Relation Extraction (SRE)). Hereby, we cast the problem of SRE into a sequence labeling problem. Two variants of CRFs will be presented in Section 2.2 to solve the two mentioned challenges.

- **Extraction Layer:** During inference, we use the well-known Viterbi algorithm [152] in order to label unannotated text data. Afterwards, the labeled text is converted to facts. To follow the *Linked Data principles* [23] (see also Section 1.1), facts can be normalized in *Text2SemRel* to exploit the full power of the gained structured data. Using already existing URIs is the preferred way to publish Linked Data. Thus, if a controlled vocabulary for entities of the modeled domain of interest is already available, this vocabulary can be used as reference for normalization. A simple sliding-window heuristic is currently used to solve this step (see Section 2.2.5).

- **Data-Storage & Representation Layer:** The extracted facts are encoded in form of subject, predicate and object triples, using the Resource Description Framework (RDF). In our case, subjects and objects are typed entities and the predicate encodes a relation that holds between the involved entities. As an example the fact, that the expression behavior of gene ITGB4 is altered in thyroid carcinoma is encoded with the triple:
  <`http://bio2rdf.org/geneid:3691`,
  `http://www.dbs.ifi.lmu.de/~bundschu/AlteredExpression`,
  `http://bio2rdf.org/mesh:D009362`>
  Optionally, additional information such as the publication id or the publication itself can be stored as well. This is solved by means of Reification. See Section 4.3 for more details.

  Text2SemRel comes with a graph-based, interactive visualization framework, which allows simple keyword queries over the extracted ER-graph (see Section 2.3).

2.1.4 Related Work

Information Extraction

The ultimate goal of information extraction (IE) is the automatic transfer of unstructured textual information into a structured form. However, the way from an unstructured, noisy source to a structured representation form of high quality is challenging. IE comprises very different tasks, ranging from named entity extraction, relationship extraction and the extraction of structures such as tables or lists. The type of the used text varies in granularity (e. g. sentence level vs. document level) and heterogeneity (e. g. template machine generated documents vs. unstructured documents - domain specific sources vs. open domains such as the web). Naturally, applied methods can be classified as well (hand-coded vs. learning-based - rule-based vs. statistical). And finally the output of the information extraction system varies in terms of structure. One option is to simply annotate the unstructured documents with all identified mentions, while the other main option is to write the identified mentions to a database. In the latter case, in order to assemble the database with high quality information, a normalization step is usually needed. [168] provides a recent survey of the broad field of IE and aligns existing work with the just outlined dimensions. We will discuss related work on Relation Extraction (RE) in more depth, since *Text2SemRel* is based on a novel RE algorithm.

Named Entity Recognition NER can be seen as the process of finding mentions of named objects in running text. Most popular are named entities such as people, locations and organizations. Various competitions have been carried out in this context such as the Automatic Content Extraction (ACE) [71], the Message Understanding Conference (MUC) [90] or the Conference on Computational Natural Language Learning (CoNLL) [74]. A popular example in the biomedical domain is the BioCreAtIvE conference [100]. Simple dictionary approaches are usually not sufficient and suffer from low recall [61].Traditionally, this task has been tackled with rule-based based systems [42, 110] or statistical methods such as Hidden Markov Models (HMMs) [152], Conditional Random Fields (CRFs) [179] and Support Vector Machines (SVMs) [41].

Usually, named entities are derived on the sentence level, where each sentence first is split into tokens. Very often, the task of NER is seen as a sequential labeling task. Various competitions revealed that if a sufficient set of labeled sentences are available for training and if the training corpus well represents the test domain, then statistical methods are usually superior to rule-based systems.

Relation Extraction Relation Extraction (RE) classically deals with the problem of finding associations between entities within a text phrase (i. e. usually, but not necessarily, a sentence). Solving this problem is essential, when trying to build ER graphs from text. The problem of extracting relations from free text is a particular difficult one, since an algorithm has to reason over noisy, non-local clues over diverse semantic and/or syntactic structures in a sentence. Numerous approaches differ in goals and characteristics, which makes it hard to classify them strictly. This section aligns existing work along several dimensions:

Initial Input *Text2SemRel* like other systems (e. g. [173, 107]) rely on a hand-labeled training corpus, where the relevant entities and relations are marked manually. Another approach is to rely on a handful of user-provided seed facts or target relations and then to bootstrap new facts [32, 1]. This bootstrapping is achieved by finding textual robust patterns , which might express the desired relation and at the same time generalize well. [1] includes a confidence measure to judge the quality of the extracted facts.

Mostly, the systems that have different initial input also have a different notion of the task of extracting relations. On the one side, the methods which rely on a hand-labeled training corpus consider the relationship extraction problem usually as follows: first, the entities are extracted from text and second, for a given fixed pair of entities, the type of relationship that holds between the pair is extracted. As we will see in Section 2.2, *Text2SemRel* differs in that, because our systems treats RE as sequence labeling task. On the other side, the systems that bootstrap facts from an initial seed of target relations usually define the problem as extracting all instances of entity pairs for which a certain type of relation holds.

Granularity of Relations The granularity of relations to extract varies among systems. A common form of relation, the *type* or *instance of* relation, can be solved with standard NER systems (e. g. [192, 74]), since lists of certain type or lists of entities must be returned (e. g. all proteins in a text collection). However, most systems typically focus on a small number of entity types. In contrast, the PANKOW system [56] is a pattern-based solution to find arbitrary *type* relations. The corresponding paradigm is named the self-annotating web. The system starts with the extraction of all proper nouns from a web page. Afterwards, a set of hypothesis phrases, pattern-based phrases that might encode the correct *type* relation, are send to the Google Web Service API to retrieve the number of hits for each hypothesis phrase. Finally, the highest ranked instance-concept pair from the result set is chosen. C-PANKOW [57] represents an extension of PANKOW that comes with an improved pattern generation process. Text2Onto is a complete framework for ontology learning and includes algorithms for learning type, subclass, part-of, and semantic relations [58].

In this thesis, we are interested in extracting semantic relations, or typed relations (such as *born_in*). Throughout this thesis we use the terms relation extraction (RE) and semantic relation extraction (SRE) as synonyms to refer to the combined task of detecting and characterizing a relation between two entities. This task is sometimes also referred to as binary relation classification in the literature [168]. Popular examples of systems that tackle this task from free text are [54, 163, 111]. However, only very few systems take a step forward and provide facts from free-text in canonical form.

Techniques The simplest way to extract relations from textual data is to analyze co-occurrence statistics of entities. Entities which often co-occur are likely to encode a relation [155]. However, these techniques are not designed for extracting typed relations. The main techniques for RE are either rule-based, pattern-based or learning-based.

As in NER, rule-based systems are a valid option to extract typed relations. In [171, 133, 110], first order rules around the context of two occurring entities are extracted. Some systems also combine rules with machine learning and learn rule weights [198].

Pattern-based algorithms have a long tradition and one of the founders of Google itself, Sergey Brin, has worked on this topic. He is the inventor of the famous DIPRE (Dual Iterative Pattern Relation Expansion) patterns [32]. The DIPRE algorithm works as follows: One seed example of a relation is given to the system. DIPRE searches the internet for further instances of the seed example. The surrounding context of the instances is used to create regular expressions. Once all regular expressions are build for the seed example, a combination of wild-card expressions and the found regular expressions searches for new seed instances. Much research builds on this famous pattern based algorithm. A particular popular example is the Snowball system [1] that builds on the same idea but builds patterns in a slightly different way. In addition, a strategy for evaluating the quality of the patterns and of the found quintuples is introduced. In the same line of pattern-based algorithms is the Espresso system [148]. [177] expands pattern-based algorithms by including deep linguistic information. Most pattern-based algorithms start with a small set of sample seeds and bootstrap the patterns iteratively. The bootstrapping of patterns is usually controlled by different motivated quality measures. A systematic comparison along with the introduction of the Pronto system can be found in [29].

In general, machine learning methods play a huge role for RE. Here the most popular types of methods are either feature-based or kernel-based. In the first case, a flat set of features is extracted from the input and a classifier such as a SVM, a decision-tree or a maximum-entropy classifier is used. A systematic investigation of the feature space for such methods can be found in [111]. Other examples of feature-based methods are e. g. [87, 197].

While these feature-based algorithms treat RE as classification task, *Text2SemRel's* RE algorithm is based on sequence labeling, thus we use an algorithm that models sequences. In sequence labeling tasks, CRFs have outperformed HMMs in many real-world applications such as NER or gene prediction tasks [179]. Other approaches that treat RE as sequence labeling problem are [15, 64]. However, [64] restricts to label relations only, no entities can be extracted with the here proposed model. In addition, they do no take into account a conceptual schema behind the extraction process. Furthermore, the model proposed in [64] is restricted to work only for bibliographic texts and not for free-text in general. While *Text2SemRel* aims at targeted extraction, the system of [15] aims at Open Information Extraction (see next paragraph), a very ambitious, relation-independent extraction paradigm. Recently, [15] report an update of *TextRunner*, in which they introduce a new RE component that is also based on Conditional Random Fields. The CRF-based RE component used in *TextRunner* has been published slightly after the here proposed contribution for Relation Extraction [34]. The CRF-based RE component in *TextRunner* nearly doubled precision as well as recall in comparison with the former RE component based on a naive Bayes classifier. *TextRunner* cannot store facts in canonical form. In addition, *TextRunner* cannot judge whether the extracted fact encodes for a meaningful real-world relation. As a concrete example, *TextRunner* extracts the fact *[Prague, wrote, Metamorphosis]* from following sentence: *'Prague is where Kafka wrote the Metamorphosis'*. *Text2SemRel* does not extract this fact, since according to a meaningful conceptual schema, the relation *writes_act* cannot hold between an entity of the city class and an entity of the act class (see Section 2.2.1 for more details).

The main idea of kernel methods for RE [194, 40, 39, 195] is to represent sentences as trees or graphs and to to design kernels, which capture the similarity between these structures. The most common structures for kernel-based methods in the IE community are parse trees or dependency graphs.

Type of Domain RE has been studied in the context of various domains, including news articles [71], emails [110], social web projects such as Wikipedia [113] or the biomedical domain. A particular challenging goal is to extract relations from an open-end domain such as the web [14]. The latter ambition represents a new paradigm for IE, named Open Information Extraction[76]. Open Information Extraction is a relation-independent paradigm that is tailored to very heterogeneous domains such as the web. A system implementing this paradigm is the already mentioned *TextRunner*. Note that open information extraction systems may not yield high precision at reasonable recall when compared to traditional RE frameworks [15]. *Text2SemRel*, in contrast, aims at optimizing both precision and recall but of course this accuracy comes with a prize: our system is quite general, scales well to large textual collections from a certain domain and can be adapted to other domains as well (see Section 2.1.3), but it does not scale to the web. Also it should be noted that in many domains, such as the biomedical, it is not acceptable for biomedical researchers (i. e. the end-users) to accept neither low recall nor low precision.

Another example of an open information extraction system, which relies on probabilistic graphical models is *StatSnowball* [198]. This system was also published later than the work presented here [34].

In general, the difficulty to extract relations from web sources varies tremendously dependent of the domain. While open information extraction from the web is quite challenging, RE from Wikipedia is rather simple, because Wikipedia infoboxes, tables which represent an almost structured form, can be utilized [113]. In contrast, extracting relations from emails is quite demanding, since emails are often written informally. This affects pre-processing steps such as Part-Of-Speech (POS) tagging or phrase chunking, which in turn hurts the performance of the overall system. Since *Text2SemRel* is evaluated in the biomedical domain, we will review work in this domain more carefully. Extracting facts from the biomedical literature is challenging, especially if we want to represent facts in canonical form, since normalization of biomedical entities such as proteins or genes to URIs is not trivial [99].

In the biomedical context, RE has most often been applied to identifying relations between proteins [24, 165, 155, 38, 162]. [63] focus on detecting associations between proteins and subcellular locations, whereas [159] extract relations between genes, drugs and cell-lines in the context of cancer. Approaches for extracting relations between genes and diseases are less prominent [158, 55], however this area is attracting increasing attention. The different approaches vary in the granularity of the relation extraction process itself. While most studies focus only on detecting relations, a small number of approaches also attempt to extract and characterize the type of relation between entities [158, 163, 162]. For example, [193] set up an interactive system where NLP methods are applied to generate a set of candidate relationship features, which are evaluated by biological experts to generate a final set of relationship features. [158] set up a system called *SemGen*, which attempts to characterize the semantics of the relations based on whether a gene causes, predisposes, or is simply associated with a disease. In this system, gene entities are identified using existing NER taggers [159, 181]. Disease entities are identified with the help of MetaMap [6], a program that maps biomedical text to concepts in the UMLS Metathesaurus [30]. In a subsequent step, each gene-disease pair is classified into one of the relational categories with the help of manually inspected indicator rules. On a test corpus of 1000 sentences a precision of 76% is reported. [127] propose a heuristic post-processing strategy for *SemGen* that aims at selecting the semantic relations that are most likely to be correct. Recently, [54] proposed a method to retrieve genes related to prostate cancer by identifying six gene-prostate cancer relations. Almost all of the related work here assumes one important prerequisite for Relation Extraction given,i. e. , that the NER step has already been performed. Entities in this domain are typically short phrases representing a specific object such as 'TP53'. Even tough NER in the biomedical domain is in general not easy, the recognition of well-defined objects, such as genes, proteins, drugs and diseases, has achieved a sufficient level of maturity such that it can form the basis for the next step: Relation Extraction. The first critical assessments for relation extraction algorithms have already been carried out (see e. g. the BioCreAtIvE II protein-protein interaction benchmark[2] or the TREC 2007 Genomics benchmark[3]). Whereas most early research focused on the mere *detection* of relations, the classification of the *type* of relation is of growing importance [158, 54, 163].

Fact Gathering Systems Recently, systems that automatically build huge knowledge bases with semantic facts have been successfully applied. Among the most popular examples are YAGO (Yet Another Great Ontology) [176], DBPedia [11], and the Kylin/KOG project [188]. The focus of these systems is the extraction of facts from semi-structured or almost structured data sources such as Wikipedia. Consequently, the accuracy of extraction in this domain is of high quality. Only very few systems aim to extract canonical facts from free-text sources. Besides *Text2SemRel* we highlight SOFIE [178], the Kylin project [188]. Popular examples of large-scale IE systems that extract non-canonical facts are DIPRE [32], Snowball [1], StatSnowball [198] and TextRunner [15]. Famous manual approaches for creating large-scale knowledge bases are e. g. WordNet [80] or the famous Cyc project [120] from Cycorp, Inc.[4].

The representation of the facts and the usability of these systems play also a crucial role. *Text2SemRel* approaches this problem with an easy to use graph based visualization framework. To make the resulting graph easy to query, we use a keyword-based query approach. Systems, which built on top of such extracted knowledge bases often rely on complex query languages to retrieve important subgraphs. One problem with these approaches is that end-users such as biomedical researchers are often not able and/or willing to query these systems in a SQL like language.

2.1.5 Contributions and Outline

Our proposed method *Text2SemRel* provides a novel framework for constructing entity-relationship graphs from text. The main contributions in this chapter are:

1. A system which extracts facts from text. *Text2SemRel* can convert facts into canonical form, provided controlled vocabularies of entities are available. The resulting facts can be directly used in a formal ontology. Being compliant with the Linked Data Principles (see Section 1.1) by using existing URI's and RDF as representation format, the extracted knowledge base can be easily integrated in existing knowledge bases.

2. A novel model for Relation Extraction based on Conditional Random Fields.

[2] http://biocreative.sourceforge.net/biocreative_2_ppi.html
[3] http://ir.ohsu.edu/genomics/
[4] http://www.cyc.com/

3. A comparison of the novel developed RE method with existing state of the art methods for RE.

4. In contrast to other existing systems, *Text2SemRel* does not rely on a complicated query language. Instead, the system comes with an easy to use, graph-based visualization framework. The system supports interactive exploration with simple keyword search over the Entity-Relationship graph.

The rest of the chapter is organized as follows: In Section 2.2 we describe the novel method for RE based on the framework of CRFs. Here, we describe the developed variants of the model, give details about the implementation and finally describe the fact extraction module (see Section 2.2.5). In Section 2.4, we validate the proposed method experimentally on two data sets. The first data set is freely available and provided by the BioText[5] group from Berkeley university. The second data set is an in-house generated data set for extracting semantic gene-disease relations from the GeneRIF database. In Chapter 4 we prove the large-scale applicability of *Text2SemRel* by applying the newly developed RE algorithm to the whole GeneRIF database. The resulting ER graph is subject of further study with respect to the quality of the extracted network (see Chapter 4).

2.2 A Novel Approach for Knowledge Base Construction with Conditional Random Fields

Our goal is to develop a method that automatically identifies entities from text and extracts typed relations among them. In this section, we are presenting our developed approach to tackle this problem. In particular, we are looking for an approach, which is able to incorporate a rich set of features to capture the richness of textual data. There will be the need to incorporate non-local, contextual clues to decide whether or not a relation is encoded in the text. Therefore our method must be able to easily handle this type of complex features. Conditional Random Fields [119], probabilistic models for segmenting and labeling sequence data, are due to their discriminative nature able to easily model arbitrary, long-range and highly dependent features. CRFs have been applied with much success to the task of NER. It would be highly desirable to find a way to cast the task of SRE into a sequence labeling problem and as a consequence could make use of a method which has been proven to be highly competitive for this task. In particular, we will introduce two novel variants of CRFs that solve entity recognition and relation extraction in a combined fashion. The first variant can be applied to encyclopaedic-style articles. Under encyclopaedic-style articles we understand documents that discuss a specific entity or concept. In this setting, we will propose a variant that labels entities and relations in one step. The second variant represents an important extension, which is able to extract entities and their relations from general free-text and not only encyclopaedic-style articles. We will also show that this model significantly outperforms the first variant on encyclopaedic-style

[5] http://biotext.berkeley.edu

data collections. In what follows, we will first motivate how we can treat the task of SRE as sequence labeling task. In contrast to most previous work, we do not assume that the entities are given in advance.

2.2.1 Named Entity Recognition and Semantic Relation Extraction as Sequence Labeling Task

Entities and their relations are extracted from sentences. A sentence in turn can be seen as a consecutive *sequence* of words, whereby the words or tokens are the elements of the sequence. The task in many information extraction problems is to assign single labels to each of the elements in the sequence. E. g. in the case of NER, we try to assign labels of entity classes to word tokens. Sequential labeling tasks, also known as *sequential supervised learning problems* [69], can be formulated as follows: Let $(\mathbf{x}, \mathbf{y})$ denote a pair of sequences where the tokens $x_1, x_2, \cdots, x_n$ are words and $y_1, y_2, \cdots, y_n$ are token labels that indicate the class membership of the corresponding input tokens. The complete training set consists of M such sequence pairs. The goal is to build a function f that correctly predicts a new label sequence $\mathbf{y} = f(\mathbf{x})$ given the input sequence $\mathbf{x}$.

The most easy way to solve this problem is to predict each label independently. This transfers the problem to a collection of multi-class classification tasks with each label representing a different class. Any classifier of choice can be utilized for this task. However, usually the labels in the sequence labeling setting are not independent. For instance, in POS tagging or grammatical tagging, where the task is to assign the lexical class to a word, it is almost impossible to have a verb immediately following a determiner. Treating these structured prediction problems independently often hurts system performance, since no dependencies on the labels is taken into account. So, what we usually want to take into account is local sequence information to improve the performance. This leads us to a second option to solve a sequence labeling problem: predicting *globally* the output sequence y. CRFs are one such example of models, where training and testing is performed over the whole sequence (see next Section 2.2.2). However before we can train a CRF, we have to find a proper way to incorporate the information about the relations between entities into the sequence.

Labeling Entities and Relations in a Textual Sequence

Prior to the learning of the fact model based on CRFs, *Text2SemRel* needs a labeled training corpus (see Figure 2.1, pre-processing layer). Once the sentences are labeled with entities and relations, our aim is to learn features that are able to generalize for new unseen example sentences. Given the entity classes and relations of interest in form of a semantic data model or conceptual schema, there are several ways to encode entities and relations in a textual sequence. This step is important to provide the learning algorithm with the best possible representation.

Before humans can start to label a corpus, we have to split a sentence into its tokens.

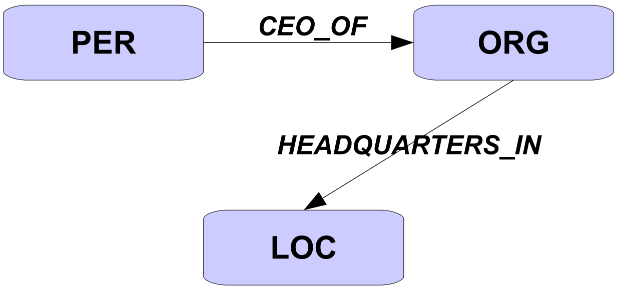


Figure 2.2: **Example of a sonceptual schema.**

Text2SemRel relies on the powerful tokenization capabilities from Lingpipe[6]. Lingpipe provides sentence models for biomedical text as well as sentence models for standard English texts such as news wire articles.

The Conceptual Schema behind *Text2SemRel* relies on a description about entities and their relationships in the real world. This description might come in form of an Entity-Relationship model [51] or another form of formal description such as an ontology (only the type-level representation is needed). Another option is to choose only a subset of entity classes and their relations from a larger schema. Figure 2.2 shows a simple example of a conceptual schema consisting of three entity classes and two relations. The representation of entity and relation classes is needed at two points in our approach. First, it ensures that we do not introduce wrong or noisy facts during the learning phase. In particular, the schema ensures that no wrong state transitions are introduced in the fact learning model (see e. g. Figure 2.4). Second, the conceptual schema is used as background knowledge during inference once a model has been learned. This guarantees that no facts are extracted that violate the conceptual schema (see Section 2.2.5). Once the type-level representation is defined, human annotators label instances of entities and relations occurring in a text collection in the desired domain. Thus, the next question to answer is, given a semantic data model or conceptual schema, how do we encode the entity classes and their relationships in the text sequence, such that it is best suited for the learning model?

Encoding Entity Classes in Text Instances of a specific entity class that are mentioned in the text sequence are assigned the label of the entity class. Here we are facing several design issues of how we want to present training examples to *Text2SemRel*. The easiest way of doing this is to assign the label of the entity class to every token that belongs to the entity. Every token which does not belong to an entity class of the predefined semantic data model is marked with a special label *O* standing for *Outside* (IO coding scheme). Let's assume that there is only one entity class in the conceptual schema. Then we have three possible state transitions in a sequence, the first from outside to outside and the

[6] http://alias-i.com/lingpipe/

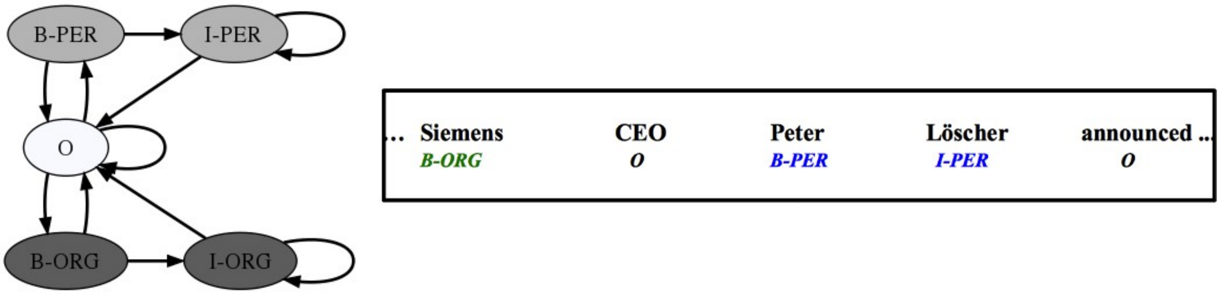
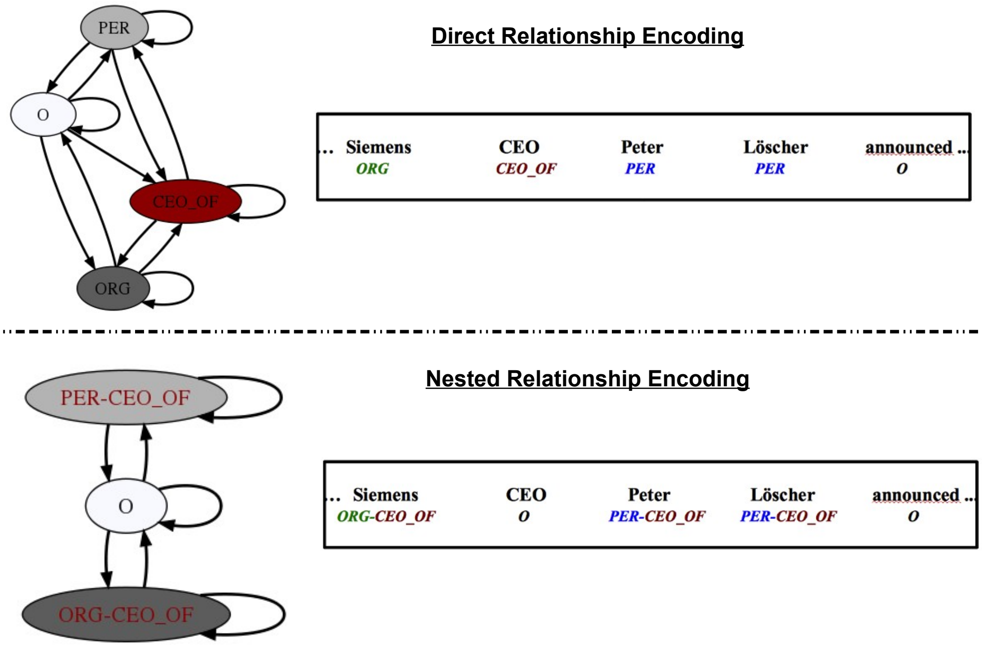


Figure 2.3: **Sequence labeling example for entity mention recognition using BIO coding scheme.** Allowed state transitions are shown on the left.

remaining two: from outside an entity class to inside an entity class and vice versa. We can further refine the transitions by introducing new flags. One such well-known refinement is the BIO coding scheme [156] originating from the Natural Language Processing (NLP) community. Here, we mark in the text sequence whether we are at the beginning of an entity instance, inside an instance or outside. Certain state transitions are constrained by default. E. g. , the transition from inside an entity class to the beginning of the same entity class is excluded by definition. *Text2SemRel* is currently using the BIO scheme. A recent comparison of different coding schemes and their influence on performance in named entity recognition systems can be found in [157]. Figure 2.3 shows an example sentence encoded with the BIO scheme, where two entity classes occur (*PER* standing for a person class and *ORG* standing for an organization class).

Encoding the relationships between entities in the text *Text2SemRel's* scope is targeted information extraction, i. e. the system is aiming at maximizing precision and recall, coming with the cost of acquiring labeled training data. More important, *Text2SemRel's* aim is to extract concise facts in the form of triples. As we will show in this paragraph, this scope affects the way how relations are encoded in a textual sequence.

The most intuitive way of annotating a relationship in text is to simple assign the relationship label to the tokens in the sentence that are indicative for the relation. Note that this style of labeling relations in a sequence has been proposed in the work of [15] for open information extraction. In the following we will call this way of annotation *Direct Relationship Encoding* (see Figure 2.4). However, despite its intuitive way of labeling, this approach comes with difficulties. First, deciding which part of a sentence encodes for a relationship between entities is not straightforward and a highly subjective task. For instance, a relationship might be expressed not only by a single word but also by several words. Very often contextual cues will be required to determine if a word token is indicative for a relationship. In addition, the indicative words do not have to be consecutive tokens in the sequence. If the relation is indeed encoded by consecutive tokens, two different human annotators might not agree about the start and/or the end of the indicative tokens. The measure of agreement of two or several human annotators is the so-called inter-annotator agreement and it serves as indicator of how difficult the annotation task is. Note that



Direct Relationship Encoding

.. Siemens	CEO	Peter	Löscher	announced ..
ORG	CEO_OF	PER	PER	O

Nested Relationship Encoding

.. Siemens	CEO	Peter	Löscher	announced ..
ORG-CEO_OF	O	PER-CEO_OF	PER-CEO_OF	O

Figure 2.4: **Direct relationship encoding vs. nested relationship encoding.** Resulting state transitions are shown on the left. For the sake of simplicity we only show the IO scheme here.

the inter-annotator agreement represents usually the upper-bound for a learning algorithm with respect to the maximal possible performance. Thus, it would be preferable to find a way to annotate the relations in the text without affecting the inter-annotator agreement too much. This would be very important in order to not bias the fact learning algorithm.

Even more important, our aim is to extract entities and relations for which there are already class definitions in form of a conceptual schema. Therefore, there is no need to label the indicative words.

Instead, an alternative way of annotating relations in a textual sequence is proposed. Instead of labeling directly the indicative tokens for the relationship, we encode the relation at the positions of the entity instances. This coding schema is named *Nested Relationship Encoding* (see Figure 2.4). Therefore, a fact is encoded in the label sequence by two participating entities that are connected by the same type of relation. Note that the entity classes have to be maintained with respect to the domain and range of the relation. I. e. no combination between entity classes and relations are allowed that are not valid against the conceptual schema.

In principle, the here discussed ways of encoding relationships (direct vs. nested) are interchangeable and it could be determined in advance, which kind of style should be used in *Text2SemRel*. This will depend on the nature of the relations in a text collection and accordingly on the conceptual schema behind the task. If there are many different types of relations between two entity classes, direct relationship encoding will confuse the learning model. However, currently we use nested relationship encoding anyway. In Section 2.4.3

we will present an application for deriving semantic gene-disease relations. Here it turns out that the two discussed ways of annotating relationships indeed differ and the second proposed way is less susceptible against annotation errors.

Encyclopaedic-style Articles vs. General Free Text The type of text has a direct influence on the labeling. As we will see, the extraction of facts from document collections such as Wikipedia is easier than from general free text. In what follows, we will describe the different labeling styles with respect to the underlying text collection.

- *Encyclopaedic-style Articles:* These data collections represent a rich source for fact extraction and as additional advantage, entities and relations can be extracted more accurately. Popular examples of these document collections are *Wikipedia*[7], *WikiGenes*[8] or *GeneRIFs* from the EntrezGene[9] database. The special form of these data collections simplifies the task of extracting facts, because of the appealing property that the articles in these collection already refer to a specific entity a priori (e. g. the Wikipedia entry of Angela Merkel[10] refers to herself). In this setting, we can define the task of extracting facts as finding all relations (if any) that hold between the a priori given entity (also called *key entity*) and all other entity mentions in the text. This formulation clearly simplifies the task, since the subject of the triples we want to extract, is already given. As a consequence, the key entity is not labeled in the text, only the other entities and their relations will be encoded in the label sequence. Here, a fact is encoded by the given key entity plus a subsequence of $\mathbf{y}$ that, at each position of the subsequence, consists of a combination of the relation label and the entity class label. As we will see in Section 2.2.3, this task formulation will allow us to solve the NER step and the SRE step jointly in one step.

  Nested relationship encoding is used. Besides the given relations that can hold between two entity classes (according to the conceptual schema), we introduce an additional label that expresses the semantics of a *negative association*. This is done for each pair of entity classes that have at least one relation according to the conceptual schema. We do this for each such pair exactly once, even though several relations between two entity classes may hold. As an alternative *Text2SemRel* could also introduce for each relation between two entity classes a kind of typed negative association for this specific relation. This decision will depend on the nature of relations in the text under investigation.

- *General free text:* We refer to general free text as text, where we cannot make such an assumption just made for entities a priori. No subject is given in advance, which complicates the task of fact extraction. In contrast to the case of encyclopaedic-style articles, we will have to train several CRF models for fact extraction, when extracting

[7] http://www.wikipedia.org/

[8] http://www.wikigenes.org/

[9] http://www.ncbi.nlm.nih.gov/gene/

[10] http://en.wikipedia.org/wiki/Angela_Merkel

facts from general free text. In particular, as we will in Section 2.2.3, we train a CRF model for each pair of entity classes occurring in a conceptual schema. For each pair of entity classes in a conceptual schema, we propose following labeling procedure:

1. As done for encyclopaedic-style articles, a label is introduced that expresses the semantics of a *negative association*. This is done for each pair of entity classes exactly once, even though several relations between two entity classes may hold.
2. If a sentence consists of a single fact, than we generate a new training pair ($\mathbf{x}$,$\mathbf{y}$) using nested relationship encoding.
3. Sentences are complex and sometimes several facts will be encoded in one single sentence. Recall that a fact is encoded in conjunction by two participating entities that are connected by the same type of relation. We set up the following labeling rule: For each true fact in a sentence, a new training instance ($\mathbf{x}$,$\mathbf{y}$) is generated using nested relationship encoding. The other entities, which are not involved in the fact under consideration are assigned their entity label plus their corresponding type of relation.
4. Each negative fact (such as a negation) is encoded with a nested relationship and the introduced negative association.

The need for a human in the loop. Obviously, we need humans to label examples for relationships between entities, since the possibilities to state a relation in free text are enormous and there are no thesauri or ontologies, which store this information. But for named entities such as genes, diseases or people there are controlled thesauri, which consist of named entities and their synonyms. E. g. , there are several controlled vocabularies for disease terms such as the Medical Subject Headings (MeSH) [123] or the Unified Medical Language System (UMLS) [30]. Why do we then need humans to label the entities? Or in other words, why do we need an intelligent entity extraction algorithm at all? This is exactly one of the lessons learned from the numerous competitions in the area of NER. As mentioned earlier, dictionary approaches may suffer from low recall. Usually authors that publish their work do not take into account controlled vocabularies, when they formulate their findings. In addition, the used thesaurus might not be specific enough and cover only broader terms. E. g. , in the MeSH thesaurus there is a concept for *breast cancer*, but no concept for *stage I breast cancer*, a more fine-grained variant of the disease. However, it is important to be able to recognize these specific terms as well. Other examples, where simple dictionary matching approaches fail are coordination (e. g. *breast and colon cancer*) or recognizing false positives (e. g. *breast cancer 2 gene*). To summarize, in order to achieve best possible recall and precision, we need an extraction algorithm, which is able to deal with the just outlined difficulties. In addition, when entities are labeled by humans with relevant domain knowledge, we can capture the knowledge which is of interest for the user.

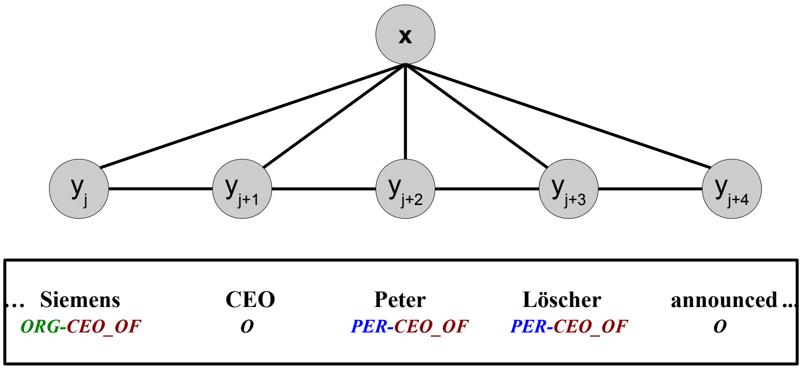


Figure 2.5: **Example undirected graphical model for a textual sequence.**

2.2.2 Conditional Random Fields

CRFs are a special form of Undirected Graphical Models or Markov Random Fields (MRFs) which are in turn a special class of Probabilistic Graphical Models (PGMs). PGMs merge two important fields of applied mathematics: Graph Theory and Probability Theory. PGMs can be seen as diagrammatic representations of a probability distribution [22], coming with some nice properties for the design and analysis of machine learning algorithms. E. g. , conditional independence properties of a given model can be obtained by visual analysis of the graph. A PGM consists of nodes and edges. The nodes are random variables, while the edges represent probabilistic dependencies between them. In MRFs, the edges of the graph are undirected, while in Directed Graphical Models, as the name implies, the edges are directed. The complete graph in a PGM gives us a notion about how the joint distribution of the model can be decomposed. The joint distribution over the node variables $p(\mathbf{n})$ of a MRF is defined by

$$p(\mathbf{n}) = \frac{1}{Z} \prod_{\mathbf{C}} \psi_C(\mathbf{n_C}), \tag{2.1}$$

where $\mathbf{C}$ is the set of maximal cliques of the underlying graph and $\psi_C(\mathbf{n_C})$ are strictly positive, but arbitrary, real-valued functions (often called potential functions). Thus, the joint distribution can be decomposed into a product of strictly positive functions defined over sets of maximal cliques that are local to the graph. A clique in an undirected graph is a subset of the complete node set, such that for every two nodes in the subset there is an edge connecting the two nodes. In other words, a clique is a fully connected subgraph. A maximal clique is a clique that cannot be extended by adding an additional node. The consequence that potential functions act on maximal cliques is that conditionally independent random variables do not appear in the same maximal clique. Since $\psi_C(\mathbf{n_C})$ can be an arbitrary function, a normalization factor Z is introduced, which sums over all possible variables of $\mathbf{n}$. This ensures that the distribution $p(\mathbf{n})$ is normalized properly:

$$Z = \sum_{\mathbf{n}} \prod_{\mathbf{C}} \psi_C(\mathbf{n_C}). \tag{2.2}$$

The reader is referred to [22] for a more theoretical derivation of the factorization properties in MRFs.

CRFs are discriminative models, which model the probability of an output sequence $\mathbf{y}$ given the input sequence $\mathbf{x}$. This modeling choice allows us to easily include rich, overlapping features, since only independence assumptions about $\mathbf{y}$ are made but not about $\mathbf{x}$. By using the product rule of probability and from Equation 2.1, we can derive a general CRF from a Markov Random Field [115]:

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})} = \frac{p(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{y}'} p(\mathbf{y}',\mathbf{x})} = \frac{\frac{1}{Z}\prod_{\mathbf{C}} \psi_C(\mathbf{x_C},\mathbf{y_C})}{\frac{1}{Z}\sum_{y'}\prod_{\mathbf{C}} \psi_C(\mathbf{x_C},\mathbf{y}'_{\mathbf{C}})} = \frac{1}{Z(\mathbf{x})}\prod_{\mathbf{C}} \psi_C(\mathbf{x_C},\mathbf{y_C}). \quad (2.3)$$

Coming back to the here considered task of extracting entities and relations from a sentence, a common design modeling choice is to model the sentence with a linear-chain CRF, where y_i is conditionally independent given its predecessor y_{i-1} and the whole input sequence $\mathbf{x}$ (see Figure 2.5). Based on Equation 2.3 and the resulting graphical structure of a linear-chain CRF (see Figure 2.5), the conditional probability $p(\mathbf{y}|\mathbf{x})$ becomes

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\prod_{i=1}^{N} \psi_i(y_i, y_{i-1}, \mathbf{x}) \quad (2.4)$$

with N being the length of the input sequence and $Z(\mathbf{x})$ being an instance-specific normalization factor. In the following we restrict to the case of linear-chain CRFs. Recall that the potential functions have to be strictly positive. They usually take the form

$$\psi_i = \exp(\sum_{k=1}^{K} \lambda_k f_k(y_i, y_{i-1}, \mathbf{x})), \quad (2.5)$$

while the use of the exponential function ensures that the potential functions are strictly positive. $f_k(y_{j-1}, y_j, \mathbf{x})$ is an arbitrary feature function, K the number of feature functions and λ_k is a weight for each feature function that can range from $-\infty$ to ∞. Each feature function f_k represents the strength of interaction between subsequent labels, dependent on the input sequence. The corresponding feature weight λ_k specifies whether the association should be favored or disfavored: Higher values of λ_k make their corresponding label transitions more likely. λ_k should be negative, if the feature tends to be off for the correct labeling and around zero in case the feature is uninformative. We use only binary feature functions of the form:

$$f_k(y_{j-1}, y_j, \mathbf{x}, j) = \begin{cases} 1 & \text{if WORD= Loescher} \in x, \text{label}_{-1}(y) = \text{B-PER}, \text{label}_0(y) = \text{I-PER}; \\ 0 & \text{otherwise} \end{cases}$$

In the above example, a transition feature is shown, which essentially is a feature that is based on an adjacent pair of labels. One might also consider to define features, which only depend on a single label to provide a form of redundancy in case for rarely occurring combinations of label pairs. The task for parameter learning is now to estimate the model parameters λ_k for each feature function f_k.

Parameter Estimation

Learning the parameters θ for CRFs is expensive and thus is an active field of research (see e. g. [186, 70]). Therefore, we only present the principled ideas how to estimate the parameters λ_k for a linear-chain CRF. Given a set of training data $\mathcal{T}$ i. i. d. given in form of M sequence pairs $\{\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\}_{m=1}^{M}$, the parameters λ_k can be estimated by maximizing the log-likelihood[11]

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{m=1}^{M} \log p(y^{(m)})|x^{(m)}) && (2.6)\\ &= \sum_{m=1}^{M} \log\left(\frac{1}{Z(\mathbf{x}^{(m)})}\exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1}^{(m)}, y_i^{(m)}, \mathbf{x}^{(m)})\right)\right) && (2.7)\\ &= \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1}^{(m)}, y_i^{(m)}, \mathbf{x}^{(m)}) - \sum_{m=1}^{M} log(Z(\mathbf{x}^{(m)})) && (2.8)\end{aligned}$$

A common choice to avoid overfitting is the inclusion of a regularization or smoothing parameter $-\sum_{k=1}^{K}\frac{\lambda_i^2}{2\sigma^2}$ [52]. The smaller the variance term σ, the smaller is the chance that a single large weight dominates a decision. With the introduction of the regularization parameter, the task becomes to maximize the penalized log-likelihood. After we have substituted the model of a linear-chain CRF into the penalized likelihood and taking the partial derivates with respect to λ_k, we get:

$$\frac{\partial\mathcal{L}(\theta)}{\partial\lambda_k} = \tilde{E}[f_i] - E[f_i] - \frac{\lambda_i}{\sigma^2}. \qquad (2.9)$$

The reader is referred to [179] for a more detailed derivation. $\tilde{E}[f_i]$ is the empirical feature expectation, which is calculated by simply counting the number of times each feature occurs in the training data. $E[f_i]$ is the model expectation of feature f_i and originates from the derivate of the normalization factor $\sum_{m=1}^{M} log(Z(\mathbf{x}^{(m)}))$. Computing $E[f_i]$ involves the summation over all possible label sequences, which is impractical. However, the Markovian structure of the CRF allows the use an efficient dynamic programming algorithm. More specifically, the famous forward-backward algorithm [152] to compute expectations over label sequences can be used. Once the model expectation is computed, the gradient of the objective function can be computed. In its simplest form, the gradient ascent method can be used. However, it turns out that it converges too slow and the use of approximate methods is needed [179].

2.2.3 CRF Model Design

In this section, the two developed CRF variants are described, which extract entities and relations. In both variants, we restrict to use only local dependencies between labels to keep

[11] Note that the product over potential functions from Equation 2.4 can be written as a sum into the exponential function.

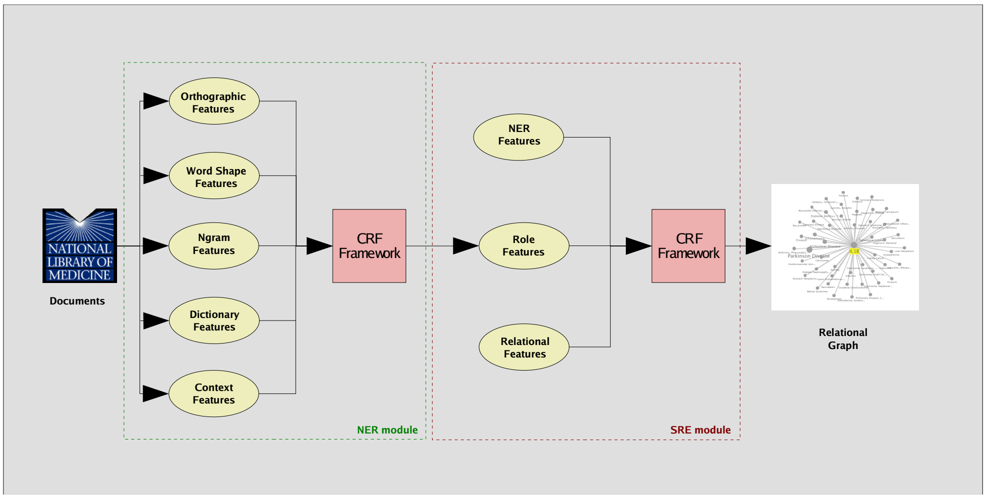


Figure 2.6: **Cascaded CRF workflow for the combined task of NER and SRE**. In the first module, a NER tagger is trained with the above shown features. The extracted role feature is used to train a SRE model, together with standard NER features (optional) and relational features.

inference tractable and cast the problem of identifying relations into a sequence labeling problem. Since the relations are encoded at the positions, where the entities occur in the sequence, it is suitable to restrict to linear-chain CRFs which have been applied to the task of NER with much success. As a consequence we can make use of a method that has been proven to be highly competitive for this task. Due to their discriminative nature, CRFs can easily incorporate arbitrary, non-local features from the input sequence (see Section next 2.2.4). This characteristic will be very suitable for tackling the task of relation extraction.

The first variant, the *one-step CRF* is more suitable for encyclopaedic-style articles (see Section 2.2.1), in return it is able to perform entity recognition and relation extraction in one single step. The second variant, the *cascaded CRF* consists of two single CRF models, the first model for identifying entities and the second model for identifying entities and their relations. The output of the first CRF model is provided as input to the subsequent CRF. Finally, we discuss how the type of text influences the model design.

Cascaded CRF for NER+SRE

In this setting, we treat the problem with a classical pipeline approach, where the output of a classifier trained for one specific problem is used as input to the next classifier, which solves the subsequent problem. Instead of learning jointly the classifier pipeline with global inference as is done in previous work for the task of noun phrase chunking [180], we restrict to cascaded training and rely on a simple but effective two-stage model. In the cascaded setting, two CRFs are trained: a CRF for NER and a second CRF for solving the combined task of NER+SRE. The trained CRF for NER is first applied to identify all entity mentions in the textual sequence. In addition to standard local features, the identified entity mentions are then used as additional input features to help solve the NER+SRE problem (Figure 2.6). The entities identified in the first step serve as soft constraints in the second model. There is no hard rule that after the second step, the entity mentions have to be exactly the same as after the first NER step. However, this entity feature will enforce a very strong correspondence between entities identified in the first step and between the final predictions.

One-step CRF for NER+SRE

Here we only consider text collections that refer to a *key entity* such as e. g. Wikipedia, Wikigenes or GeneRIF phrases. Thus, the key entity is already given. All other entities in the text phrase, so-called *secondary entities*, are assumed to stand in some relation to the key entity. In the cascaded setting described above, SRE is treated with a classical pipeline approach. The special nature of encyclopaedic-style document collections allows us to solve NER and SRE in one step. This is also reflected by the labeling, as a secondary entity's label encodes the type of the entity plus the type of relation with the key entity (see Section 2.2.1). Note that the key entity itself has not to be explicitly mentioned in the text. To illustrate the assumption made in this setting, we give an example from the biomedical domain. GeneRIF sentences represent a similar style of text in the biomedical domain such as Wikipedia. GeneRIFs describe the function of a gene/protein, the key entity, as a concise phrase. Consider the following GeneRIF sentence linked to the gene COX-2:
'Expression in this gene is significantly more common in endometrial adenocarcinoma and ovarian serous cystadenocarcinoma, but not in cervical squamous carcinoma, compared with normal tissue.'
This sentence states three disease relations with COX-2 (the key entity), namely two altered expression relations (the expression of COX-2 relates to endometrial adenocarcinoma and ovarian serous cystadenocarcinoma) and one unrelated relation (cervical squamous carcinoma).

Encyclopaedic-style Articles vs. General Free Text.

Several CRF models are trained, when extracting facts from general free text in order to alleviate ambiguities. In particular, we train one such model for each pair of entity classes in the conceptual schema for which there exists at least one single defined relation. E. g. , given the conceptual schema from Figure 2.2, results in training two CRF models for SRE. One model for the pair between the entity classes *PER* and *ORG* and the second for the entity classes *ORG* and *LOC*. It is important to note that even though several relations between two entity classes may hold, only single model is trained for all relations holding between this pair of entity classes. For the NER task, a single global model is trained.

In encyclopaedic-style article collections, we do not have such difficulties and train only one global model for SRE independent of the conceptual schema.

2.2.4 Feature Classes

The choice of features is an essential part in any IE system and much of a system's performance depends on the accurate definition of features. Due to their discriminative nature, CRFs are able to incorporate many rich and highly dependent features. One of the strengths of CRFs is that they soften hard rules, i. e. features get assigned probabilistic weights according to their maximum likelihood estimate. In this section, general characteristics of the features used in *Text2SemRel* are described, while in Section 2.4.1 we provide detailed information about the implementation of these features, when discussing our experiments. *Text2SemRel's* features can be roughly divided into local features, context features and features consisting of external knowledge. These can further be grouped into features designed rather for NER and features designed rather for SRE. For instance, in case of SRE it will be often necessary to reason over contextual clues located quite far away from an entity, that's why the contextual features are expected to be more helpful for SRE. Note that all features, except the external knowledge features, are derived from the training data.

Local features These are extracted from single tokens or parts of a single token. Many IE systems use a kind of standard feature set with slight derivations across systems. Local features are expected to be more meaningful for NER. *Text2SemRel* uses essentially the following:

- *Word Token*: The word tokens represent the most simple features, but might already be indicative for a specific label.

- *Orthography*: Entities often share common orthography, e. g. they are often capitalized or consist of digits. A complete list of orthographic features is shown in Table 2.1.

- *Word Shape*: This feature is a normalization, where a word is broken down into its shape. E. g. two different words that have the same length and are capitalized encode the same word shape. The word shape of the word 'Angela' is normalized to 'Xxxxxx'. A further option is to prune this normalization such that all adjacent letters are merged into a single one ('Xx' in the example above).

- *Character N-gram*: The character N-grams are defined for a specific window length and are consecutive substrings of N items from a given word sequence. Character 3-grams of 'Angela' are: 'Ang', 'nge', 'gel', 'ela'.

- *Prefix*: Prefix features are consecutive substrings of N items from a given word sequence, with the additional constraint that they must start off with the beginning of the word sequence. The prefix of the word 'kinase' with length $N = 3$ would be 'kin'.

- *Suffix*: Suffix features are consecutive substrings of N items from a given word sequence with the additional constraint that they must end with the last character of the word sequence. The suffix of the word 'kinase' with length $N = 3$ would be 'ase'.

- *Role Feature*: Special feature designed for SRE, which indicates that a specific entity class has been recognized in the input sequence. Every token from the input sequence which is predicted to be member of an entity class gets assigned this feature. This information comes from a NER system. *Text2SemRel* uses a CRF for NER.

External Knowledge Features Incorporating external knowledge into an IE system often improves performance and is an essential part of many IE systems (see e. g. [114, 157]). *Text2SemRel* uses two different types of external knowledge features: (i) dictionaries consisting of entities and (ii) keyword lists that are indicative for specific relations. While the accuracy of these dictionaries is usually very good, its coverage is typically quite low. The idea behind dictionary matching features is that if some substring of the input token sequence matches a dictionary entry, the entity class of the corresponding dictionary should get more likely. Even though there are lot of resources available on the web, dictionaries are domain-dependent features and might sometimes be difficult to obtain. Some recent work tries to automatically generate dictionaries from large collections of unlabeled text (see e. g. [114, 77]).

Context Features These features extract characteristics from the surroundings of a current word x_i. CRFs can ask arbitrary questions about the input token sequence $\mathbf{x}$, which makes it straightforward to incorporate context features in this model. Context features are the crucial factor for being able to handle the task of extracting relations with CRFs, since a relationship between two entities is most of the time expressed in the environment of one of the participating entities. Recent research indicates that binary relationships in general can often be expressed with a small number of lexico-syntactic patterns [15]. In the future *Text2SemRel* could make use of these results and use such lexico-syntactic patterns as additional features.

- *Conjunction*: The input for conjunction is an interval $[-N; N]$ and a set of other feature classes for which conjunctions shall be extracted. A conjunction feature takes features at a specific position inside the specified interval and combines the feature with position-specific information. For instance, let's assume we have a given interval $[-2; 2]$ and the set of features for which conjunction is specified, consists of the prefix feature and the suffix feature class. Let's assume the current word is 'Merkel' and the previous word is 'Angela'. For $N = -1$ we get one conjunction feature for the suffix 'ela' and position $N = -1$ as well as one conjunction feature for the prefix 'Ang' and position $N = -1$. Note that the conjunction feature is expected to be helpful for both NER and SRE.

- *Entity Neighborhood*: A special feature for SRE, that gets as input the following information: predicted substrings for which the role feature (see local features) is assumed to be true, an interval $[-N; N]$ and a set of other classes of features for which entity neighborhood shall be considered. The neighborhood feature searches in the specified interval around the predicted entity if specific classes of features can be found.

- *Window*: The window feature considers the same input space as the conjunction feature does. But in contrast to the conjunction feature, no position-specific information is extracted, only the information that a defined feature occurs in the given window.

- *Negation*: The input space for the negation feature is a set of feature classes. This feature is active, if none of the defined feature classes can be found in the input sequence.

2.2.5 Fact Extraction

In the last section we have seen how features are defined and how parameters can be learned for NER and SRE based on CRFs. This section explains how facts can be extracted from a new unseen input sequence $\mathbf{x}$ once a model has been learned. Facts in *Text2SemRel* are represented as triples consisting of two involved entities and a typed relation. In general, facts can come in canonical form and non-canonical form. Fundamental to the extraction of facts in canonical form is the normalization of entities to URIs. Facts that are in canonical form are unambiguous, i. e. the participating entities and relations are well-defined and no different identifiers exist for the same entity. On the contrary, facts that are expressed in natural language are usually non-canonical. Entities and relations appear in numerous different spelling variants, because people writing text do not take controlled vocabularies into account. Note that the extraction of non-canonical facts is already a huge benefit and is very important for several techniques such as document retrieval, summarization and knowledge discovery.

Since facts extracted from natural language are usually in ambiguous form, the facts extracted by *Text2SemRel* are, at first, ambiguous. Indeed, the relations come in normalized form, since *Text2SemRel* scope is targeted information extraction, where relations are defined in advance. Nevertheless, the recognized entity mentions are non-canonical. *Text2SemRel* can identify entities that are not in any dictionaries, for this reason the system is more flexible than an information extraction approach that uses only controlled vocabularies to identify entities. But at this stage, facts extracted by our approach are still ambiguous, i. e. the same fact can be expressed with different variants. As just outlined, the extraction of non-canonical facts is already a huge advancement and enables a lot of powerful applications, but in order to exploit the full power of the gained structured data, facts need to be normalized. We refer to the task of normalizing a mention of an entity in free text to a controlled vocabulary as *entity normalization*, a task that is related to entity resolution or record deduplication (see e. g. [18]). *Text2SemRel* provides a simple optional normalization against a controlled vocabulary that is based on a sliding window approach.

In general, it cannot be assumed that the subsequences that are labeled as part of an entity class by *Text2SemRel*, can be mapped directly to an entry of a controlled vocabulary. The trained model might only predict a part of the entity correctly or the corresponding entry of the controlled vocabulary might be a substring of the recognized entity mention. E. g. when identifying disease mentions from text, the proposed system might recognize the mention 'lethal metastatic pancreatic neoplasm' but the corresponding entry in the controlled vocabulary is 'pancreatic neoplasm'. Due to the sliding window approach, this does not necessarily mean that we are not able to establish the correct link to a controlled vocabulary.

From Sentences to Facts

We will start explaining how facts are extracted from encyclopaedic-style articles. Afterwards we will generalize the fact extraction for free text. Finally, we will explain the proposed sliding window heuristic for entity normalization in more detail.

Before we start to extract facts from plain sentences, we have to split the sentence into its tokens first. *Text2SemRel* relies on the powerful tokenization capabilities from Lingpipe[12]. Lingpipe provides sentence models for biomedical text as well as sentence models for standard English texts such as news wire articles.

Encyclopaedic-style Articles Encyclopaedic-style document collections have an appealing property: The key entity is already known in advance and the task in this setting is to predict the relation between the key entity and all other entities in the text. Due to the particular nature of these document collections a one-step CRF can be defined that is able to label a sequence with entities and relations jointly. But we can also apply the cascaded CRF here. The reader is referred to Section 2.2.3 for details about the models.

Essential for both models is how to predict for a new unlabeled token sequence $\mathbf{x}$ the most likely label sequence $\mathbf{y}^*$. This problem can be solved efficiently by using a Viterbi-style algorithm [152]. In particular the most likely label sequence can be obtained in general by maximizing

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \exp(\sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, \mathbf{x})). \tag{2.10}$$

The difference between the one-step CRF and the cascaded CRF is that the cascaded CRF takes, as additional input, features originating from the prediction of a NER-based CRF (the role features). These features are computed in advance with the Viterbi-style algorithm from Equation 2.10 utilizing a set of feature functions specifically designed for NER. However, once this feature is computed, the algorithm for predicting the most likely label sequence is the same for both models and the output is the most likely label sequence $\mathbf{y}^*$. Up to now the facts are encoded in the sequence pair $(\mathbf{x}, \mathbf{y}^*)$ and the question arises how facts can be extracted from this sequence pair?

[12] http://alias-i.com/lingpipe/

Algorithm 1: Pseudocode for fact extraction from encyclopaedic-style articles.

```
   Input: x // a sentence
   Input: $keyEntity$ // subject of fact extraction
   Data: $controlledVocList$ // thesauri for each entity class (OPTIONAL)
   Result: $facts$ // list of triples
1  $facts \leftarrow \emptyset$;
2  $\mathbf{y}^* = \arg\max_{\mathbf{y}} \exp(\sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, \mathbf{x}))$;
3  $entityMentions \leftarrow$ ExtractEntityMentions($\mathbf{x}, \mathbf{y}^*$);
4  foreach $entityMention \in entityMentions$ do
5      if IsProperFact($entityMention, keyEntity$) then
6          if $canonicalForm$ then
7              $entityClass \leftarrow$ GetEntityClass($entityMention$);
8              $controlledVoc \leftarrow$ GetControlledVoc($entityClass$);
9              $entityURI \leftarrow$
               SlidingWindowNormalization($\mathbf{x}, entityMention, controlledVoc$);
10             if $entityURI \neq \emptyset$ then
11                 $relationClassURI \leftarrow$ GetRelation($entityMention$);
12                 $facts$.Add([$keyEntityURI, relationClassURI, entityURI$]);
13             end
14         else
15             $relationClassURI \leftarrow$ GetRelation($entityMention$);
16             $facts$.Add([$keyEntityURI, relationClassURI, entityMention$]);
17         end
18     end
19 end
```

In addition to the input sequence $\mathbf{x}$, two further inputs need to be specified in advance: the key entity, which is already given as well as a set of controlled vocabularies for each entity class that are subject of extraction (provided that we want to extract facts in canonical form). The extraction of facts now works as follows (see Algorithm 1): The method `ExtractEntityMentions` starts with extracting all entity predictions made by the model (either one-step CRF or cascaded CRF). This is done by extracting all consecutive subsequences from $\mathbf{y}^*$ for which the labels are predicted to be part of an entity class (i. e. where the labels are not equals to the label O standing for Other). The string representation of the entity mentions (subsequences of $\mathbf{x}$) and the predicted entity classes are saved accordingly. For example, for the sequence pair from Figure 2.4 we would extract the string *Siemens*, the entity class *ORG* and the token index ti. The second recognized entity string would be *Peter Loescher*, the entity class *PER* as well as the token indices $ti + 2$ and $ti + 3$. The method `IsPorperFact` (Algorithm 1, line 5) checks if the entity classes connected by the relation label maintain the conceptual schema. Here the semantic data model comes into play and we consider only facts for further processing that do not violate the schema. Afterwards, if we want to extract facts in canonical form, we try to map each of the recognized entity mentions to an URI (line 7-9, Algorithm 1). Therefore, `GetControlledVoc` returns the corresponding vocabulary for the entity class of the entity mention. Afterwards we try to normalize the extracted string to an URI with the method `SlidingWindowNormalization`. If this method succeeds and returns an URI, the relation class for the predicted entity is extracted by returning the type of relation from the label sequence from the entity mention (method `GetRelation`). The triple *[keyEntityURI,relationClassURI,entityURI]* is saved. Alternatively, if canonical form extraction is not required, we simply store the fact *[keyEntityURI,relationClassURI,entityMention]*. The method `Add` takes into account domain and range of the extracted fact and changes the order accordingly if needed (e. g. given a *CEO_OF* relation, the first argument has to be a person and the second has to be of type organization).

General Free Text The property that encyclopaedic-style articles have, namely that the key entity is given a priori, does not hold for general free text anymore. This makes the fact extraction more complicated and as a consequence we resort to the cascaded CRF in this setting. The fact extraction starts with applying a trained CRF for NER (see Algorithm 2, line 2). The recognized entity mentions are used as features for the SRE based CRF. Dependent on the number of predicted entity mentions, the algorithm either starts to compute the most likely label sequence $\mathbf{y}^*$ (see Algorithm 2, line 5-7) or computes the top N most likely label sequences (Algorithm 2, line 8-10). If a sentence has more than two entity mentions, than all true facts might not be encoded in one single most likely label sequence anymore. Consider, for instance, a sentence that encodes two facts consisting of three entities and two different types of relations. In this case, you can only find one proper fact in the most likely label sequence. Thus, we investigate the top N most likely label sequences, whether or not we can find additional entity pair mentions that might encode a proper fact. Following the idea of [91], the N most likely label sequences can be extracted in a linear forward Viterbi pass and backward pass with the A* algorithm. The backward pass uses the Viterbi scores as completion results. Note that N is currently set to be equals to the number of found entity mentions. The method `GetEntityPairMentions` extract all pairs of entity mentions from the input sequence $\mathbf{x}$ by considering the predicted label sequence $\mathbf{y}^*$ or the top N predicted label sequences from **topNViterbi**. The corresponding subsequences of $\mathbf{x}$ and $\mathbf{y}$ are saved accordingly. Afterwards we check for each entity pair, whether it represents a proper fact. Again, the method `IsPorperFact` checks if the entity classes connected by the relation label maintain the conceptual schema. Again, the semantic data model comes into play and we consider only facts for further processing that do not violate the schema.

Sliding Window Approach for Entity Normalization The goal of the normalization step is to link entities mentioned in free text to URIs. This represents a key step to provide well-defined facts. In order to keep the fact extraction quality as high as possible, *Text2SemRel* currently uses a very strict normalization criterion. A potential entity

Algorithm 2: Pseudocode for fact extraction from general free text.

```
Input: $x$ // a sentence
Input: $controlledVocList$ // thesauri for each entity class (OPTIONAL)
Result: $facts$ // list of triples
$facts \leftarrow \emptyset$;
$\mathbf{y}^*_{\mathbf{NER}} = \arg\max_{\mathbf{y}_{\mathbf{NER}}} \exp(\sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_{NER_k} f_{NER_k}(y_{i-1}, y_i, \mathbf{x}))$;
$entityMentions \leftarrow$ GetNumberOfEntityMentions($\mathbf{y}^*_{\mathbf{NER}}$);
$entityPairMentions \leftarrow \emptyset$;
if $entityMentions == 2$ then
    // get most likely label sequence with role features
    $\mathbf{y}^* = \arg\max_{\mathbf{y}} \exp(\sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, \mathbf{x}))$;
    $entityPairMentions \leftarrow$ GetEntityPairMentions($\mathbf{x}, \mathbf{y}^*$);
else if $entityMentions > 2$ then
    // get most likely $N$ label sequences with role features
    topNViterbi $\leftarrow$ ComputeNBestViterbiOutput($\mathbf{x}, \mathbf{y}^*_{\mathbf{NER}}, N$);
    $entityPairMentions \leftarrow$ GetEntityPairMentions($\mathbf{x}$, topNViterbi);
end
foreach $entityPair \in entityPairMentions$ do
    if IsProperFact($entityPair$) then
        if $canonicalForm$ then
            // get URIs for first and second entity in the same way as
               done in Algorithm 1, line 6-8
            if $entityURI1 \neq \emptyset$ AND $entityURI2 \neq \emptyset$ then
                $relationClassURI \leftarrow$ GetRelation($entityMention1$);
                $facts$.Add([$entityURI1, relationClassURI, entityURI2$]);
            end
        else
            $entityMention1 \leftarrow$ GetEntity($entityPair, 1$);
            $entityMention2 \leftarrow$ GetEntity($entityPair, 2$);
            $relationClassURI \leftarrow$ GetRelation($entityMention$);
            $facts$.Add([$entityMention1, relationClassURI, entityMention2$]);
        end
    end
end
```

candidate has to be matched completely to one of the entries or its synonyms of the controlled vocabulary of the corresponding entity class. We assume a controlled vocabulary or thesaurus to consist of an URI and a set of associated string representations (e. g. synonyms or orthographic variants). We resort to a sliding window approach, that scans the immediate neighborhood of the found entity mention in the input sequence $\mathbf{x}$, since

Algorithm 3: Pseudocode for sliding window normalization.

```
   Input: x // a sentence
   Input: $entityMention$ // consecutive subsequence of $x$
   Input: $controlledVoc$ // same class like entity class of $entityMention$
   Result: $URI$ // URI representing the entity
1  $candidateURIs \leftarrow \emptyset$;
   // get string representation of predicted entity from CRF
2  $[start; end] \leftarrow$ GetPosition($entityMention, \mathbf{x}$);
3  $entityString \leftarrow$ GetStringRepresentation($\mathbf{x}, [start; end]$);
   // check if entityString is an entry of the controlled voc.
4  $candidateURI \leftarrow$ GetURI($controlledVoc, entityString$);
5  if $candidateURI \neq \emptyset$ then
       // We have the first candidate
6      $candidateURIs$.Add($candidateURI$);
7  end
   // check if we have further candidates
8  $candidateURIs \leftarrow$ ScanNeighborhood($\mathbf{x}, [start; end], controlledVoc$);
9  switch $controlledVoc.type$ do
10     case $hierarchical$
11         $URI \leftarrow$ GetDeepestCandidateURI($candidateURIs, controlledVoc$);
12     case $flat$
13         $URI \leftarrow$ GetURIWithLongestStringRep($candidateURIs, controlledVoc$);
14     endsw
15 endsw
```

the predicted entity mentions originating from *Text2SemRel* cannot always be mapped directly to a controlled vocabulary. First, *Text2SemRel* extracts entity mentions from text, written by people which do not take into account a controlled vocabulary. Second, the probabilistic model might only predict a part of the entity correctly or an entry of the controlled vocabulary might be a substring of the entity mention. The idea behind the sliding window approach is that even though in the majority of cases the potential entity candidate originating from the model might be a direct hit in the controlled vocabulary, we can improve the accuracy of the system by scanning the preceding or following tokens of the predicted entity. Assume e. g. that the model predicted the word ‘cancer’ to be of the type disease. Let’s further assume that the preceding token is ‘breast’. In this case we would like to extract the disease ‘breast cancer’ and not only ‘cancer’.

The normalization works as follows (see Algorithm 3): The input token sequence is given together with the predicted entity mention and the appropriate thesaurus consisting of entities and synonyms from the same entity class as the predicted entity. The algorithm starts with a test whether the predicted entity mention is a complete match against the thesaurus. If this is the case, we have a first candidate URI (Algorithm 3, line 2-7). If a disambiguation problem occurs, i. e. the entity mention refers to several URIs, then we neglect the entity mention. Note that a complete thesaurus match is defined to be true if the entity mention is a complete string of a thesaurus entry or of one of its synonyms. In contrast, a partial match is defined to be true if the entity mention is a substring of a thesaurus entry or one of its synonyms (e. g. a suffix or prefix).

The method `ScanNeighborhood` then scans the preceding and following tokens of the current entity mentions for further URI candidates (except the initial entity mention did not even match an entry in the controlled vocabulary partially). The entity mention boundaries in the input sequence $\mathbf{x}$ are given by $[start; end]$. Unless we are not at the beginning of a sentence (i. e. $x_i = 1$), the method prolongs the entity mention token-wise starting with an expansion of the entity mention to the left. After every token-wise prolonging, the new generated candidate is matched against the thesaurus. If a full match against an entry or its synonyms can be found, then the URI of the match is added to the list of candidate URIs. Again, if a disambiguation problem occurs, then we neglect the entity mention and do not extract a candidate URI. If only a partial match is detected then no URI can be added, but the prolonging continues until no partial match can be found anymore. The longest partial match of this left expansion becomes the new entity candidate (if there is one). The same token-wise expansion strategy just outlined is then conducted to the right with either the original entity mention or the new, expanded entity candidate (originating from the left expansion). Every emerging full match against the controlled vocabulary is added to the list of candidate URIs. We continue this expansion to the right unless we either have no partial thesaurus match anymore or we are at the end of the sentence (i. e. $x_i = N$). If we have not found any candidate URI up to now, we start to shrink the original entity mention and test if we can find any candidate URI in a substring of the original entity mention.

Up to now, we have a list of candidate URIs. The way how the final URI for the entity is extracted, depends on the nature of the controlled vocabulary (line 10-15, Algorithm 3). In case the controlled vocabulary is composed of hierarchical relationships such as hyponym relations, we chose the URI which represents the most specific entity. Otherwise, if we have a flat controlled vocabulary, we choose the URI which originates from the longest string matching against the thesaurus.

2.3 Visualization and Interactive Search

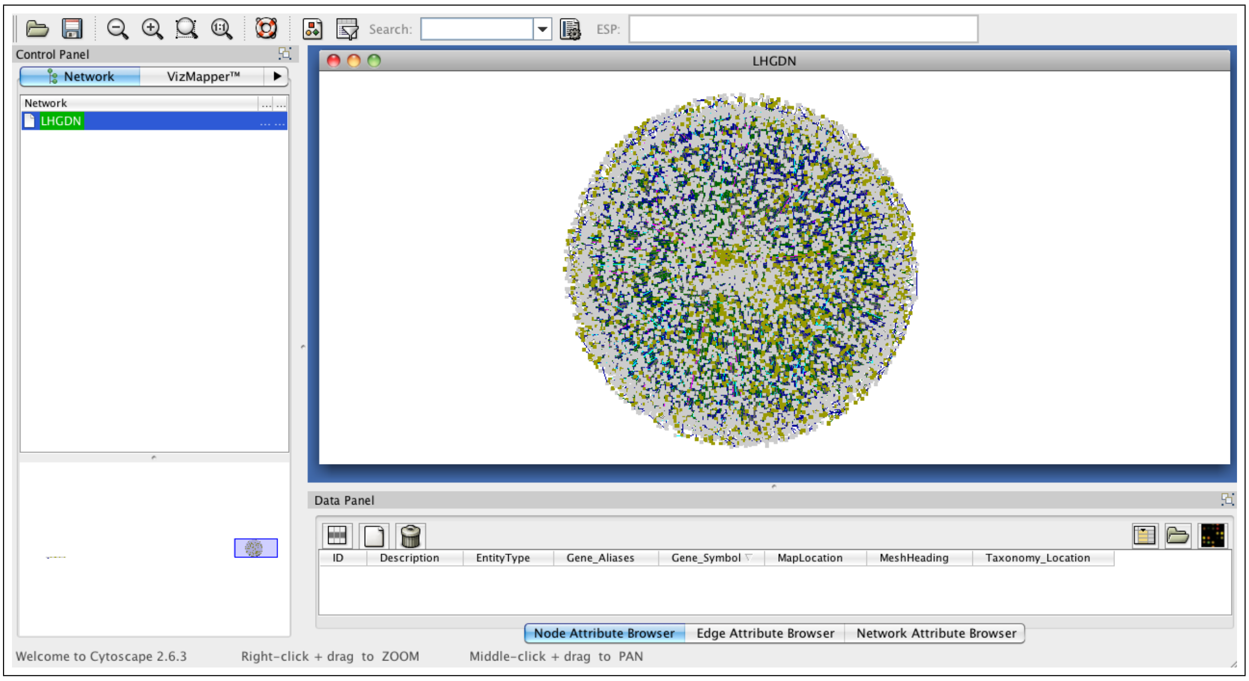
Once a knowledge base consisting of triples is constructed with *Text2SemRel*, the question arises how it can be accessed. Besides common filtering functions over the knowledge base such as filtering for specific entities and relations, *Text2SemRel* comes with an easy to use, graph-based visualization framework. The system supports interactive exploration with simple keyword search over the Entity-Relationship graph. Thus, we use a combination of paradigms originating from two different communities to access knowledge bases derived from text: (i) graph-based visualization and (ii) keyword search over structured data. The innovative combination of the above mentioned paradigms for information access in

Text2SemRel makes extracted knowledge bases easily searchable.

Graph-based Visualization This first component for information access in *Text2SemRel* arises from the field of information visualization [95]. In general, information visualization deals with the challenging task to find suitable forms of visualizations for complex information data. The motivation behind this field is that the human brain can assimilate information represented as images much faster than representations in textual form. The underlying data structure of *Text2SemRel* is an ER graph, consisting of entities such as genes, drugs, diseases etc. and of relations or facts concerning the entities. Thus, visualizing the knowledge base as a graph is intuitive and enables the user to get a fast overview. E. g. , due to the graph-based visualization it can be easily seen which entities are the most connected ones (the hubs) in the ER-graph. As another concrete example, it can be easily seen, which parts of the knowledge base are disconnected. All these things make the graph-based visualization a helpful feature for knowledge discovery.

Keyword Search over Structured Data The second used component for simplifying information access in *Text2SemRel* originates from the database community with the attempt to make relational database management systems (RDBMS) as easily searchable as keyword-based search engines (see e. g. [19, 2]). Usually, a formal query language such as SPARQL for the semantic web or SQL for traditional RDBMS is used to make structured information repositories accessible. A common criticism of these formal query languages is that it is hard and uncomfortable for end-users to pose queries. So even if all knowledge could be transformed into structured form, many users would still be unable to retrieve this information. For simple queries an intelligent user interface may automate the formulation of a certain type of query. However, for any non-standard search task the user has to write his own structured search queries which requires deep formal thinking and good knowledge of the structure of the data store.

Furthermore, despite the recent advances in the IE domain, it will not always be possible to convert all relevant information into a structured form. One reason for this is that the assessment of relevance regarding information is highly subjective and varies from end-user to end-user. Even if different end-users agree on the importance of specific entities and relations, it is still very likely that they pursue different information needs. As a concrete example, assume that two biomedical researchers are interested in facts about gene-disease relations. The first researcher focuses on gene-disease associations in Caucasian populations, while the second is interested in associations studied in the context of Japanese populations. To meet this information need, we would have to teach *Text2SemRel* to extract tertiary relations between genes, diseases and population groups. On the contrary, other researchers will have a complete different information need. It is not difficult to imagine that the number of entities and relations needed to capture all information needs will increase heavily. Therefore, methods are needed that are able to extract the most important facts (semi-)automatically, but at the same time are able to further narrow down the desired context. Therefore, *Text2SemRel* makes use of the plain text collections



Welcome to Cytoscape 2.6.3 Right-click + drag to ZOOM Middle-click + drag to PAN

Figure 2.7: **Screenshot of the representation layer.** Visualization of the whole LHGDN from Chapter 4. The complete knowledge base serves as starting point for knowledge discovery.

from which facts are extracted and annotates these with words originating from the text sources. In this way, we can additionally filter facts according to keywords and thus provide additional, powerful filtering capabilities, without the need to extract further entities and relations. Thus, the keyword-search component over the structured ER-graph will help to satisfy the manifold end-user needs.

System Description

In what follows, we describe the information access or representation layer of our proposed system in detail. Without loss of generality and for the sake of clarity we describe the information access component with the help of a concrete use case that is presented in Chapter 4, where we show how *Text2SemRel* builds a huge knowledge base consisting of semantic relations between genes and diseases.

The proposed information access layer in our system relies on the power of a network visualization tool well established in the biomedical community to visualize biological networks: Cytoscape[13] [60]. Cytoscape is a powerful tool that allows researchers to browse through huge biological networks with filtering and zooming functionality among others. We subscribe to the principled idea of using Cytoscape in combination with the literature as proposed by the Agilent Literature Search System [185]. However, our approach is different. The Agilent Literature Search System searches like traditional IR systems first

[13] http://www.cytoscape.org/features2.php

for documents based on a standard input query. Afterwards, a network of entities and relations is constructed based on the resulting document set obtained from the query.

In contrast, the starting point of our approach is the complete extracted knowledge base. Our approach first extracts all facts offline from a whole document collection. Furthermore, facts are annotated with the word tokens from the documents, where the facts were found. The same fact is often stated in publications several times, therefore a fact can be annotated with plenty of words. This will allow us to use additional powerful filtering possibilities based on keywords. In addition, we include additional information available for the nodes (i. e. the genes and diseases). This is easily achieved since we are using, according to the Linked Data principles, existing URIs from the Bio2RDF project [17]. In this way, we can include gene aliases, chromosomal locations of the genes, disease aliases, and additional descriptions without any effort. All available information for the facts, including the keywords from the publications, are indexed with the enhanced search plugin (ESP) [8], that uses the Lucene retrieval library[14] for indexing.

Figure 2.7 shows the complete literature-derived human gene-disease network (LHGDN) extracted in Chapter 4 as a starting point for further analysis. Now, a user has the following options to search for specific facts:

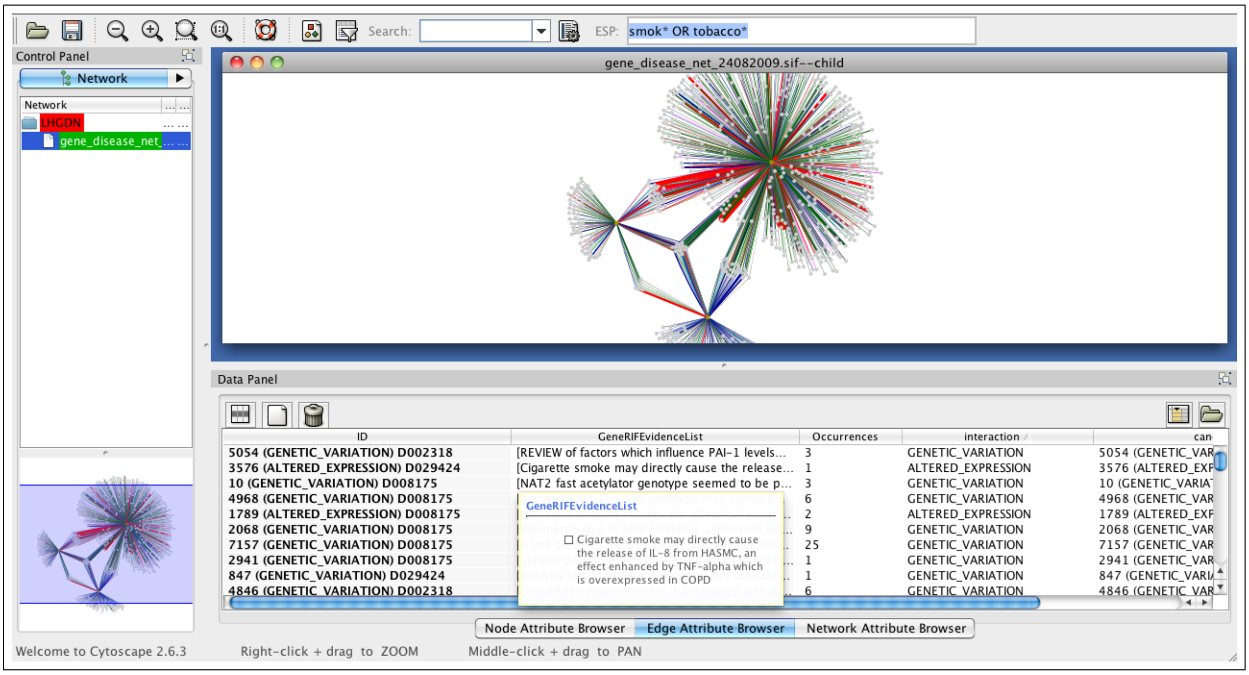
- Filtering for entities and/or types of relations and/or attributes such as in traditional RDBMS systems.

- Filtering of facts with regard to keywords. We index all available information for facts by using the ESP. The ESP allows advanced querying such as search with wildcards or range search for numbers. One of the unique features of a literature-derived network is that we can use the available unstructured information for filtering purposes. To keep sensitivity high, we only include the words of the sentence from which the fact was extracted.

- Merging of networks. Every filtering step induces a new subnetwork. These subnetworks can be merged to investigate connections between different subnetworks. Thus, we can easily isolate information but also assemble disconnected information.

As a concrete example, let's assume a researcher is interested in the following three diseases: Chronic obstructive pulmonary disease (COPD), lung cancer, and cardiovascular disease (CVD). These three diseases, sometimes referred to as 'The Big Three', are all smoking-induced diseases. 'The Big Three' account for a large fraction of mortality every year around the globe [66]. First, we would like to know more about associated genes in these diseases separately. As a first filtering step, three subnetworks are created by querying the system with the three single diseases and the option to show all associated entities, in this case genes. Afterwards, the researcher wants to know more about the connections between the diseases and thus uses the provided merge functionality. Figure 2.8 shows the result of the merged networks. Green squares are diseases, while the grey dots represent the entity type gene. The color of the edges indicate the type of relation

[14] http://lucene.apache.org/


Search:
ESP: smok* OR tobacco*
Control Panel
Network
Network
LHGDN
gene_disease_net_
gene_disease_net_24082009.sif--child
Data Panel
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.6.3
Right-click + drag to ZOOM
Middle-click + drag to PAN

ID	GeneRIFEvidenceList	Occurrences	interaction	can
5054 (GENETIC_VARIATION) D002318	[REVIEW of factors which influence PAI-1 levels...	3	GENETIC_VARIATION	5054 (GENETIC_VAR
3576 (ALTERED_EXPRESSION) D029424	[Cigarette smoke may directly cause the release...	1	ALTERED_EXPRESSION	3576 (ALTERED_EXF
10 (GENETIC_VARIATION) D008175	[NAT2 fast acetylator genotype seemed to be p...	3	GENETIC_VARIATION	10 (GENETIC_VARIAT
4968 (GENETIC_VARIATION) D008175	[illegible]	6	GENETIC_VARIATION	4968 (GENETIC_VAR
1789 (ALTERED_EXPRESSION) D008175	[illegible]	2	ALTERED_EXPRESSION	1789 (ALTERED_EXF
2068 (GENETIC_VARIATION) D008175	[illegible]	9	GENETIC_VARIATION	2068 (GENETIC_VAR
7157 (GENETIC_VARIATION) D008175	[illegible]	25	GENETIC_VARIATION	7157 (GENETIC_VAR
2941 (GENETIC_VARIATION) D008175	[illegible]	1	GENETIC_VARIATION	2941 (GENETIC_VAR
847 (GENETIC_VARIATION) D029424	[illegible]	1	GENETIC_VARIATION	847 (GENETIC_VARI
4846 (GENETIC_VARIATION) D002318	[illegible]	6	GENETIC_VARIATION	4846 (GENETIC_VAR

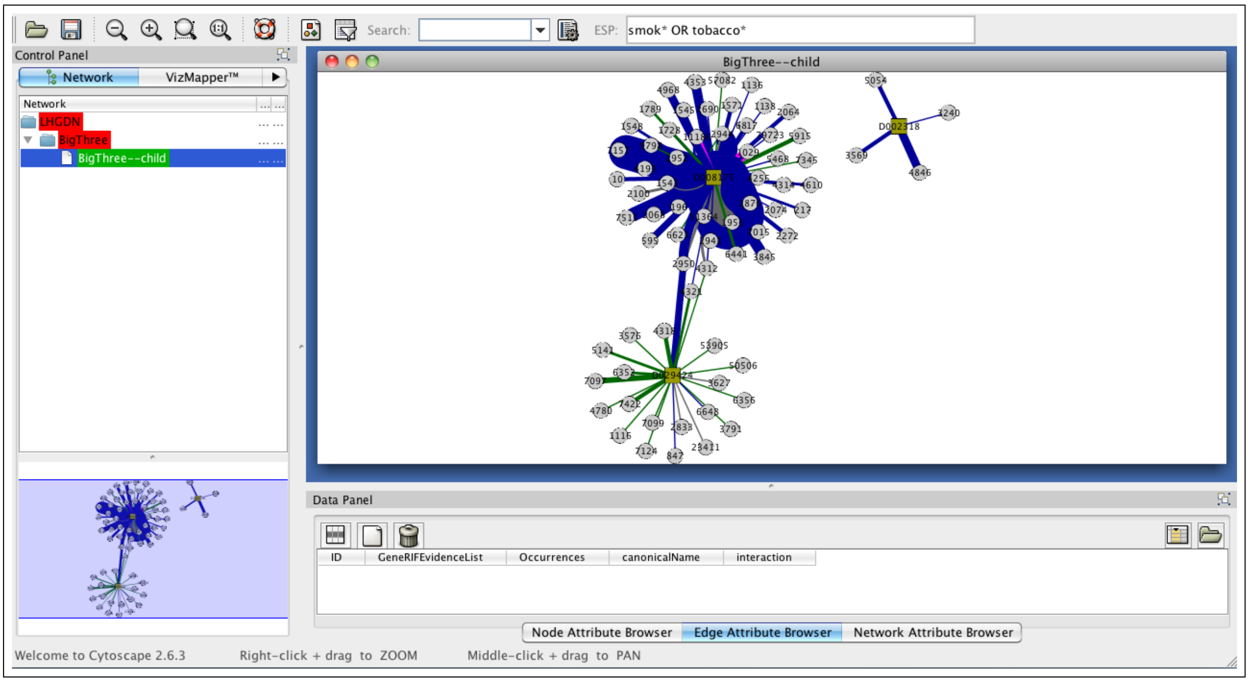
GeneRIFEvidenceList

□ Cigarette smoke may directly cause the release of IL-8 from HASMC, an effect enhanced by TNF-alpha which is overexpressed in COPD

Figure 2.8: **Screenshot of a subnetwork that shows all facts for three diseases.** The subnetwork of COPD, CVD and lung cancer are shown. Green squares are diseases, while the grey dots represent the entity type gene. The color of the edges indicates the type of relation as predicted by *Text2SemRel*. Edges highlighted in red indicate that the keyword query *smok* OR tobacco** are matched. The thickness of the edges indicate the number of times a fact was found in different publications. The data panel provides important additional context information such as the sentence where the fact was extracted from.

as predicted by *Text2SemRel*. There are five types of relations in the here shown use case: *altered expression*, *genetic variation*, *regulatory modification*, *any relation*, and *negative association*. For more details about the types of relations see Section 2.4.3 as well as Chapter 4, Section 4.3. The thickness of the edges indicate the number of times a fact was found in different publications. As a next step, we would like to filter the merged subnetwork explicitly with regards to gene-disease associations that are discussed in the context of smoke or tobacco. The keyword query *'smok* OR tobacco*'* is posed and the resulting facts, which match the criteria are highlighted in red (see Figure 2.8). The facts that match the keyword query can be filtered accordingly and a new view is created (see Figure 2.9).

During all operations the user can get additional information for the entities and facts via the data panel. The data panel functionality of Cytoscape provides views for the nodes and views for the edges. In our case, we provide gene descriptions, gene aliases, official gene symbols, and the chromosomal location of the genes. In addition, we provide the predicted type of relations and the sentence where the facts were extracted. This ensures that the user gets further insights and can decide upon the predicted fact, whether it is interesting



ESP: smok* OR tobacco*

ID	GeneRIFEvidenceList	Occurrences	canonicalName	interaction

Welcome to Cytoscape 2.6.3 Right-click + drag to ZOOM Middle-click + drag to PAN

Figure 2.9: **Screenshot of a subnetwork that shows filtered facts for three diseases.** The subnetwork of COPD, CVD and lung cancer are shown. Green squares are diseases, while the grey dots represent the entity type gene. The color of the edges indicate the type of relation as predicted by *Text2SemRel*. The keyword filter *'smok* OR tobacco*'* was applied.

or not. All the provided additional information can be used for the keyword-based filtering.

2.4 Experimental Evaluation

To assess the performance of the fact learning module, we benchmark our approach on two different relation extraction tasks. The first task is the identification of semantic relations between *diseases* and *treatments*. The publicly available data set[15] consists of manually annotated PubMed abstracts. There are seven predefined types of relations to detect: *cure, only disease, only treatment, prevents, side effect, vague, does not cure* (see Section 2.4.2). We compare our approach against the results published in [163]. [163] use a multilayer Neural Network (NN) and probabilistic generative models specifically designed for extracting relations.

The second task is the identification of semantic relations between *genes* and *diseases* from a set of concise phrases, so-called GeneRIF (Gene Reference Into Function) phrases. There are five predefined types of relations: *altered expression, genetic variation, regulatory modification, any relation, negative association*. GeneRIFs represent encyclopaedic style articles, where the key entity (in this case the gene), is already given. We compare our

[15] http://biotext.berkeley.edu/

two developed approaches with results obtained by a Support Vector Machine (SVM), and with two rule-based methods.

Compared with these state of the art approaches, we achieve very competitive results. We achieve higher or comparable accuracy on two evaluation data sets. In our experimental setting, we do not assume that the entities are given, as is often the case in previous relation extraction work. Rather the extraction of the entities is solved as a subproblem.

2.4.1 System Description

The fact learning model was implemented with the help of MALLET [132], which provides an efficient implementation for CRFs. As already mentioned earlier we restricted to the special case of linear-chain CRFs and used the default Gaussian prior provided by MALLET. In what follows, we will describe the features used in the experiments. We motivate why the chosen features make sense in the context of the biomedical domain. Furthermore, the feature descriptions are based on the entities and relations used in the experiments. A general description of various feature classes used in *Text2SemRel* can be found in Section 2.2.4. Note that the features are used in both types of CRFs (one-step and cascaded), unless explicitly stated otherwise.

Local features

- *Word Token:* The simplest features are the word tokens themselves (no stemming performed). We do not use any higher level syntactic features such as part-of-speech tags or noun phrase chunks.

- *Orthography:* Biomedical entities often yield some orthographic characteristics: They often consist of capitalized letters, include digits or are composed of combinations of both. Thus, these features are helpful in distinguishing various types of biomedical entities. These features can be easily implemented using regular expressions. The set of regular expressions used in this work is displayed in Table 2.1.

- *Word Shape:* For instance, it may be common for disease abbreviations, that digits and letters cannot appear together in the token, while for genes and proteins the co-occurrence of digits and letters is striking.

- *Character N-gram:* We also used character n-gram word features for $2 \leq n \leq 4$. These features help to recognize informative substrings like ‘ase’ or ‘homeo’, especially for words not seen in training.

- *Role Feature (only used for cascaded CRFs):* This feature indicates, for cascaded CRFs, that the first system extracted a certain entity, such as a disease or treatment entity class. This means, that the tokens that are part of an entity mention (according to the NER CRF) are labeled with the type of entity predicted for the token.

Orthographic Feature	**Regular Expression**
Init Caps	`[A-Z].*`
Init Caps Alpha	`[A-Z][a-z]*`
All Caps	`[A-Z]+`
Caps Mix	`[A-Za-z]+`
Has Digit	`.*[0-9].*`
Single Digit	`[0-9]`
Double Digit	`[0-9][0-9]`
Natural Number	`[0-9]+`
Real Number	`[-\+][[0-9]+[\.,]+[0-9].,]+`
Alpha-Numeric	`[A-Za-z0-9]+`
Roman	`[ivxdlcm]+|[IVXDLCM]+`
Has Dash	`.*-.*`
Init Dash	`-.*`
End Dash	`.*-`
Punctuation	`[,\.;:\?!-\+â]`
Greek	`(alpha|beta|...|omega)`
Has Greek	`.*\b(alpha|beta|...|omega)\b.*`
Mutation Pattern	`\w*\d+-*\D+`

Table 2.1: **Orthographic Features used in *Text2SemRel*.** Orthographic features and their corresponding regular expressions.

External Knowledge Features Since we are tackling two tasks of IE, namely NER and SRE, two classes of dictionaries are employed: (i) entity class dictionaries consisting of controlled vocabularies and (ii) relation type dictionaries, which contain indicative keywords for types of relations. Note that the presence of a certain dictionary entry in a sentence is indicative, but not imperative, for a specific entity or relation. This property is elegantly handled by the probabilistic nature of our approach. In general, a dictionary feature is active if several tokens match with at least one entry in the corresponding dictionary.

- *Entity Class Dictionaries:* The disease dictionary is based on all names and synonyms of concepts covered by the disease branch (C) of the MeSH ontology. In addition, a treatment dictionary is introduced for the disease-treatment extraction task, composed of all names and synonyms of concepts from the MeSH D branch (Chemicals and Drugs).

- *Relation Type Dictionaries:* We define four relation dictionaries for the GeneRIF data set, each composed of relation type specific keywords for the following types of relations: *altered expression*, *genetic variation*, *regulatory modification* and *unrelated.* For example, the *genetic variation* dictionary contains words like ‘mutation’ and ‘polymorphism’. For disease-treatment relations we set up dictionaries containing keywords for *prevent* and *side effect* relations. The relation specific dictionaries are provided as supplementary data (see [34]).

Contextual Features These features take into account the properties of preceding or following tokens for a current token. Context features are very important for several reasons. First, consider the case of nested entities: ‘Breast cancer 2 protein is expressed . . . ’. In this text phrase we do not want to identify a disease entity. Thus, when trying to determine the correct label for the token ’Breast’ it is very important to know that one of the following word features will be ‘protein’, indicating that ‘Breast’ refers to a gene/protein entity and not to a disease. The importance of context features not only holds for the case of nested entities but for SRE as well.

- *Word Conjunction:* Preceding or following words within an interval of $[-3; 3]$ from the current token are extracted.
- *Dictionary Window Feature:* For each of the relation type dictionaries we define an active feature, if at least one keyword from the corresponding dictionary matches a word in the window size of 20, i. e. -10 and +10 tokens away from the current token.
- *Key Entity Neighborhood Feature:* For each of the relation type dictionaries we defined a feature which is active if at least one keyword matches a word in the window of 8, i. e. -4 and +4 tokens away from one of the key entity tokens. To identify the position of the key entity we queried name, identifier and synonyms of the corresponding Entrez gene against the sentence text by case-insensitive exact string matching.
- *Start Window Feature:* For each of the relation type dictionaries we defined a feature which is active if at least one keyword matches a word in the first four tokens of a sentence. With this feature we address the fact that for many sentences important properties of a biomedical relation are mentioned at the beginning of a sentence.
- Negation Feature: This feature is active, if none of the three above mentioned special context features matched a dictionary keyword. It is very helpful to distinguish *any* relations from more fine-grained relations.

2.4.2 Disease-Treatment Relation Extraction from PubMed

Data Set

This annotated text corpus provided by [163] was generated from MEDLINE 2001 abstracts. In a total of 3570 sentences, entities describing diseases and treatments were extracted and disease-treatment relations were classified as *cure, only disease, only treatment, prevents, side effect, vague, does not cure.* Note that, in contrast to the original work, we present results for the full data set, including sentences that contain no entities at all. We believe that this setting is much more realistic than looking only at sentences

	NER			SRE	
	Recall	Precision	F-score	Accuracy (1)	Accuracy (2)
Best GM [163]	-	-	71.0	91.6	74.9
Multilayer NN [163]	-	-	-	96.6	**79.6**
cascaded CRF	69.0	75.3	**72.0**	**96.9**	79.5

Table 2.2: **Results NER+SRE for the disease-treatment corpus.** NER and SRE performance based on evaluation scores proposed by [163]. Relation classification accuracy for seven types of relations is shown for two settings: (1) when the entities are given as gold standard and (2) when the entities have to be extracted. The cascaded CRF outperforms the best graphical model approach and it shows similar performance to the multilayer neural network, where the latter approach can not be applied to the NER task, due to the large feature vectors.

where at least one of the two entities occurs. The data, enriched with supplementary annotations, are provided online[16].

Results

In this data set a key entity is not known a priori and we only report results using the cascaded CRF approach. We benchmark our approach with [163], who compared five different Graphical Models (GM) and a multilayer neural network (NN) for identifying entities and disease-treatment relations. In the first experiment, we compare the CRF for NER with the benchmark methods on the NER task. As in [163], we evaluate two settings for SRE. In the first setting, entities are assumed to be correctly labeled by hand in a pre-processing step and only the existence and the type of the relation between entities needs to be predicted. In the second setting, the entities need to be identified as well. To achieve comparable results we use identical accuracy measures, namely precision, recall and F-measure for NER, and accuracy for SRE. Precision, recall and F-measure are estimated on a token level with the MUC evaluation score[17]. We used 5-fold cross-validation, in accordance with the 80%/20% training/test split used by [163].

Table 2.2 shows the results for NER and SRE. We achieve an F-measure of 72% on NER identification of disease and treatment entities, whereas the best graphical model achieves an F-measure of 71%. The multilayer NN can not address the NER task, as it is unable to work with the high-dimensional NER feature vectors [163]. Our results on SRE are also very competitive. When the entity labeling is known a priori, our cascaded CRF achieved 96.9% accuracy compared to 96.6% (multilayer NN) and 91.6% (best GM). When the entity labels are assumed to be unknown, our model achieves an accuracy of 79.5% compared to 79.6% (multilayer NN) and 74.9% (best GM).

In summary, our cascaded CRF is clearly superior to the best graphical model of [163] in both tasks. The performance on SRE is comparable to the multilayer NN, note however

[16] http://biotext.berkeley.edu/data.html

[17] http://www.itl.nist.gov/iaui/894.02/related_projects/muc /muc_sw/muc_sw_manual.html

that this method is unable to to be applied to NER.

Discussion

The performance of the cascaded CRF on the data set provided by [163] is on par with the multilayer NN and superior to the best GM. This may be due to the discriminative nature of CRFs and NNs, which could be an advantage over the generative GM. Moreover, it should be stated that the multilayer NN does not scale well with the number of features, thus limiting its applicability [163]. In [163] the NN could not be applied to the NER task, due to the large feature vectors. Our approach can be applied to both tasks, NER and SRE, achieving very competitive results. In contrast to [163], we do not make any use of syntactic higher-level features, such as Part-Of-Speech (POS) tags or Noun Phrase (NP) chunks. When the entities are already given for the SRE task, our approach achieves very accurate results, with an increase in accuracy of 18 percentage points, compared to the case where the entities were hidden and had to be recognized as well. Consequently, the most potential for further improvement lies in the correct identification of treatment and disease entities, since accuracy significantly decreases when the entities need to be identified and were not given a priori. This is especially true for treatment entities, where performance of identifying treatments is only 64.85% (F-measure), compared to disease NER performance of 77.20% (F-measure). Thus, most errors in SRE do occur when, e. g. , in the NER step a treatment entity was missed, resulting in a consecutive error of the following SRE step. Since the definition of treatments is in general vague, possible improvements could be achieved with the inclusion of a larger and/or more refined treatment dictionary. Currently, all entries of the D MeSH branch are simply used to fill the treatment dictionary, while [163] stress the careful inclusion of subbranches of the MeSH ontology.

2.4.3 Gene-Disease Relation Extraction from GeneRIF Phrases

Data Set

GeneRIFs [135] are phrases which refer to a particular gene in the Entrez Gene database [126] and describe its function in a concise phrase. Our data set consists of 5720 GeneRIF sentences retrieved from 453 randomly selected Entrez Gene database entries and the task is to extract and characterize relations between genes and diseases in those sentences. Note that the gene entities themselves are known and do not need to be extracted. See Section 2.2.5 for more details about extracting relations in encyclopaedic-style document collection. We consider relations describing a wide variety of molecular conditions, ranging from genetic to transcriptional and phosphorylation events:

- **Altered expression:** A sentence states that the altered expression level of a gene or protein is associated with a certain disease or disease state. Example: ‘Low expression of BRCA1 was associated with colorectal cancer.’
- **Genetic variation:** A sentence states that a mutational event is reported to be

related to a disease. Example: ‘Inactivating TP53 mutations were found in 55% of lethal metastatic pancreatic neoplasms.’

- **Regulatory modification:** A sentence associates a disease to a methylation or phosphorylation. Example: ‘E-cadherin and p16INK4a are commonly methylated in non-small cell lung cancer.’
- **Any:** A sentence states a relation between a gene/protein and a disease, without any further information regarding the gene’s state. Example: ‘E-cadherin has a role in preventing peritoneal dissemination in gastric cancer.’
- **Unrelated:** A sentence claims independence between a certain state of a gene/protein and a certain disease. Example: ‘Variations in TP53 and BAX alleles are unrelated to the development of pemphigus foliaceus.’

From a biological perspective, methylation and phosphorylation events could be represented as two separate types. However, due to the lack of available examples, we considered both to be of the same type.

Two human experts with biological backgrounds annotated the corpus with an inter-annotator agreement estimated of about 84%. A more detailed data set description as well as our annotation guidelines are provided as supplementary data [34]. The annotation guidelines are available online[18]. As we did not confine the study to a specific disease model, the labeled disease entities are diverse in terms of the type, ranging from rare syndromes to well studied diseases, primarily cancer and neuro-degenerative diseases like Alzheimer or Parkinson. An entity corresponds usually to a phrase such as ‘pancreatic neoplasms’. In our work disease entities were labeled in a way that preserves as much information as possible. For example, tokens specifying the disease like ‘*lethal metastatic* pancreatic neoplasms’, were considered to be part of one disease entity. The data set was tagged with nested relationship encoding (see Section 2.2.1), since first the relations are given anyway and second the human annotators were less susceptible to annotation errors. We tried the two ways of annotation on a small subsample of 100 sentences. When using direct relationship encoding (see Section 2.2.1), the inter-annotator agreement dropped by 6%. As a concrete example, consider the following sentence:
‘A novel mutation (Leu705Val) within the Abeta sequence is reported in a family with recurrent intracerebral hemorrhages.’
While annotating the entities (genes and diseases) in the above example is relatively straightforward, the annotation of the indicative words for the relationship (genetic variation) is difficult. In the first annotation variant, it is not clear e. g. whether the annotator should annotate the token *Leu705Val* to be indicative for the relationship or not.

Results

For the second data set a more stringent criterion for evaluating NER and SRE performance is used. As noted earlier, [163] use the MUC evaluation scoring scheme for estimating the

[18] http://www.biomedcentral.com/1471-2105/9/207/additional/

	Recall	**Precision**	**F-score**
Dictionary + naive rule-based	43.31	42.98	43.10
CRF + naive rule-based	67.62	71.88	69.68
CRF + SVM (linear)	72.42	75.12	73.74
one-step CRF	73.36	78.66	75.90
cascaded CRF	76.61	79.46	78.00
CRF+ SVM (F-score + RBF)	76.63	79.48	78.03

Table 2.3: **Results NER+SRE for the gene-disease corpus**. The cascaded CRF is on par with the *CRF+SVM (F-score + RBF)* model, where the latter one requires an expensive preceding feature selection step.

NER F-score. The MUC scoring scheme for NER works at the token level, meaning that a label correctly assigned to a specific token is seen as a true positive (TP), except for those tokens that belong to no entity class. SRE performance is measured using accuracy. In contrast to [163], we assess NER as well as SRE performance with an entity level based F-measure evaluation scheme, similar to the scoring scheme of the bio-entity recognition task at BioNLP/NLPBA[19] from 2004. Thus, a TP in our setting is a label sequence for that entity, which exactly matches the label sequence for this entity from the gold standard. As a concrete example, consider the following sentence:
'BRCA2 is mutated in stage II breast cancer.'
According to our labeling guidelines (see supplementary data from [34]), the human annotators label *stage II breast cancer* as a disease related via a genetic variation. Assume our system would only recognize *breast cancer* as a disease entity (which is the corresponding MeSH taxonomy entry), but would categorize the relation to gene 'BRCA2' correctly as *genetic variation*. Consequently, our system would obtain one false negative (FN) for not recognizing the whole label sequence as well as one false positive (FP). In general, this is clearly a very hard matching criterion. In many situations a more lenient criterion of correctness could be appropriate (see [183] for a detailed analysis and discussion about various matching criteria for sequence labeling tasks). To assess the performance we use a 10-fold cross-validation and report recall, precision and F-measure averaged over all cross-validation splits. Table 2.3 shows a comparison of the baseline methods with the one-step CRF and the cascaded CRF. The first two methods (*Dictionary+naive rule-based* and *CRF+naive rule-based*) are overly simplistic but can give an impression of the difficulty of the task. Recall, that in this data set NER reduces to the problem of extracting the disease since the gene entity is identical to the Entrez Gene ID. In the first baseline model (*Dictionary+naive rule-based*), the disease labeling is done via a dictionary longest matching approach, where disease labels are assigned according to the longest token sequence which matches an entry in the disease dictionary. The second baseline model (*CRF+naive rule-*

[19] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html

based) uses a CRF for disease labeling. The SRE step, referred to as *naive rule-based*, for both baseline models works as follows: After the NER step, a longest matching approach is performed based on the four relation type dictionaries (see **Methods**). Given that exactly one dictionary match was found in a GeneRIF sentence, each identified disease entity in a GeneRIF sentence is assigned with the relation type of the corresponding dictionary. When several matches from different relation dictionaries are found, the disease entity is assigned the relation type which is closest to the entity. When no match can be found, entities are assigned the relation type *any*. The third benchmark method is a two-step approach, where the disease NER step is performed by a CRF tagger and the classification of the relation is done via a multi-class SVM. We try two different configurations. While, the first configuration is simply a SVM with a linear-kernel (*CRF+SVM (linear)*), the second configuration uses feature selection in combination with a RBF-kernel (*CRF+SVM (F-score + RBF)*). The feature vector for the SVM consists of the same relational features defined for the CRF (Dictionary Window Feature, Key Entity Neighborhood Feature, Start of Sentence, Negation Feature etc.) and the stemmed words of the GeneRIF sentences. As can be seen, the *CRF+SVM* approach is greatly improved by feature selection and parameter optimization, as described by [53], using the LIBSVM package[20]. In contrast to the *CRF+SVM (F-score + RBF)* approach, the cascaded CRF and the one-step CRF easily handle the large number of features (75956) without suffering a loss of accuracy. In the combined NER-SRE measure (Table 2.3), the one-step CRF is inferior (F-measure difference of 2.13) when compared to the best performing benchmark approach (*CRF+SVM (F-score+RBF)*). This is explained by the inferior performance on the NER task in the one-step CRF. The one-step CRF achieves only a pure NER performance of 84.27%, while in the *CRF+SVM* setting, the CRF achieves 86.97% for NER. As shown in table 2.3, the cascaded CRF is on par with the *CRF+SVM* benchmark model. Table 2.4 lists the relation-specific performance for the cascaded CRF. Recall from the beginning of this section, that we use an entity-based F-measure to evaluate our results on this data set. Clearly, there is a strong correlation between the number of labeled examples in the training data (see supplementary data [34]) and the performance on the various relations. For *any*, *altered expression* as well as *genetic variation* relations we exceed the 80% F-measure boundary. Only for two types of relations does accuracy fall below this boundary, namely for *unrelated* and *regulatory modification* relations. This moderate performance can be explained by the relatively low number of available training sentences for these two classes. In general, the CRF model allows for the inclusion of a variety of arbitrary, non-independent input features ranging from simple orthographic to more complex relational features. To estimate the impact of individual features on the overall performance for the combined NER+SRE score, we trained several one-step CRFs on the same data (one specific cross-validation split), but with different feature settings. In particular, we are interested in the impact of the various relational features. Since the relational feature setting between the two applied types of CRFs was similar, we restrict this evaluation to the one-step model here. Table 2.5 lists the impact of different features for the one-step CRF model in terms of recall,

[20] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

	Recall	**Precision**	**F-score**
Any	79.46	78.45	78.95
Unrelated	60.26	70.59	65.02
Altered expression	77.96	79.90	78.91
Genetic variation	77.76	82.45	80.04
Regulatory modification	69.17	73.28	71.16
Overall	76.61	79.46	78.00

Table 2.4: **Results NER+SRE according to types of relations.** NER+SRE performance of the cascaded CRF approach for the five different relation types according to recall, precision and F-measure averaged over the 10 cross-validation test runs.

precision and F-measure. The baseline one-step CRF setting uses features typical for NER tasks, such as orthographic, word shape, n-gram and simple context features. Since we are addressing a relation extraction task, the results are poor, as expected (F-measure 38.48 and 39.65 before and after adding dictionary features, respectively). With the advent of longer/special relational features for the relation task, our system gains a large performance increase (F-measure 67.38 after adding the dictionary window feature). The inclusion of the start window feature (F-measure increase of 4.56) and the key entity neighborhood feature (F-measure increase 2.04) both gain an additionally performance increase. The inclusion of the negation window feature moderately improves recall for the *any* relation and improves precision for *altered expression*, *genetic variation* and *regulatory modification.*

Baseline CRF	●	●	●	●	●	●
Dictionaries		●	●	●	●	●
Dictionary Window			●	●	●	●
Start Window				●	●	●
Key Entity Neighborhood					●	●
Negation Window						●
Recall	35.89	38.13	64.30	70.01	71.81	72.16
Precision	41.47	41.30	70.78	74.00	75.87	78.56
F-score	38.48	39.65	67.38	71.94	73.98	75.22

Table 2.5: **Evaluation of System Components.** Contribution of different features to the overall performance of the one-step CRF for the 9th cross-validation run. The baseline model includes orthographic, word shape, n-gram and the basic context feature.

Discussion

On the GeneRIF data set the cascaded CRF performs as well as the *CRF+SVM* model. However it should be noted that training of the cascaded CRF is much faster (factor of ten in our setting), since no time-consuming feature selection is needed. The one-step CRF cannot cope with the above mentioned methods, primarily as a result of a lower recall in the NER step. An investigation of different feature weights revealed a stronger dominance of relational features in the one-step CRF compared to the cascaded CRF. Thus, the absence of certain relational features hurts the NER performance of the one-step CRF, because the relational features are a strong indicator of an occurring disease entity in this model. The fact that for *any* relations, where our relational features are usually switched off, the performance decrease is highest (F-measure difference 1.7, compared to the cascaded CRF) supports this hypothesis. For the remaining types of relations, the one-step model can cope with the benchmark approach.

Major improvements for both approaches can be achieved with a more accurate detection of entity boundaries. The overall system performance significantly increases when relaxing the hard matching criterion to softer ones (as presented in [183]). This implies that many entity boundaries are not identified properly. On the one side, this could be partly due to labeling inconsistencies of the human annotators. On the other side, it might originate from the labeling guidelines of diseases. All variable descriptions of a particular disease, such as the form '*non-small* cell lung cancer' or '*stage I-III* endometrial cancer' had to be identified, as well as directly adjacent prepositional phrases like 'cancer *of the lung*'. This makes the task clearly more challenging. The F-measure for a soft matching criterion, when only a part of an entity has to be detected properly, increases to 85.20% (F-measure) (NER+SRE). Another performance increase can be obtained with a more accurate detection of *unrelated* relations. In our framework an *unrelated* relation is a gene-disease pair for which a phrase states that the two entities are not related to each other under a specific setting.

In contrast to previous studies, where *unrelated* relations are most often skipped, we decided to categorize them, since our corpus contains about 7% unrelated statements, which is roughly three times higher than in the work of [158]. However, for a supervised learning approach this is still a very sparse training set, resulting in a low accuracy. The same problem holds for *regulatory modification* relations, where the poor performance is again likely due to the small amount of available examples in our corpus (only 3.5% of the total number of relations). Thus, for both types of relations we expect a significant increase in performance with the inclusion of more training data.

Regarding the definition of the gene-disease relation types, we emphasize that they do not account for the etiological property underlying a specific gene-disease relation. Thus, whether or not a gene is causing the disease or is just associated with the disease pathogenesis is not encoded in the gene-disease relationships defined here. However, our predefined types and the gene-disease relations extracted on that basis can provide helpful information for further biomedical research (e. g. annotation of experiments or providing additional information for experiment design). For the identification of biomarker candidates, the information on which level of the biological dogma (e. g. DNA, RNA, protein etc.) molecules are discriminative for a certain disease, provides highly valuable information, independent of their role in the disease etiology [10]. Nevertheless, we plan to extend our relation types towards etiological information as proposed by [158].

2.5 Conclusions

Summary In this chapter we presented *Text2SemRel*, a new system to build (semi-)automatically knowledge bases from textual data consisting of facts about entities using typed relations. Facts can be converted into canonical form provided that controlled vocabularies of entities are available. Since *Text2SemRel* relies on a hand-build training corpus it is more suitable for targeted relation extraction. We consider relation extraction as sequence labeling task and an essential part of the system is the learning module for relation extraction that is based on Conditional Random Fields. We extend the framework of Conditional Random Fields towards the annotation of semantic relations from text and apply it to the biomedical domain. Our approach is based on a rich set of textual features as well as on external knowledge features. The proposed approach achieves a performance that is competitive to leading approaches. The model is quite general and can be extended to handle arbitrary entities and relation types. *Text2SemRel* comes with an easy to use, graph-based visualization framework. The system supports interactive exploration with simple keyword search over the Entity-Relationship graph and provides rich functionalities to filter information.

Discussion The power and promise of unsupervised relation extraction systems is indisputable, but these systems may not yield high precision at reasonable recall [15]. In many domains such as the biomedical, it is not acceptable for end-users to suffer neither from low recall nor from low precision. *Text2SemRel*, in contrast, aims at optimizing both precision and recall but of course this accuracy comes with a prize: our system is quite general, scales well to large textual collections from a certain domain and can be adapted to other domains as well (see Section 2.1.3), but it does not scale to the web.

Text2SemRel can convert facts into a canonical form provided that controlled vocabularies of entities are available. This is a very important prerequisite to make the assembled knowledge available for semantic-based search engines. Currently, *Text2SemRel* was tested on entities that are naturally not very polysemous. In these cases the entity normalization strategy works fine. However, a more advanced normalization strategy will be needed for entities that are highly polysemous (e. g. person names). Thus, another issue of future work will be to improve the entity normalization module.

The extraction performance of *Text2SemRel* will vary, as in other IE systems, from relation to relation. As a concrete example, extracting facts about protein-protein interactions from text is particular difficult, because many facts are often mentioned in one single sentence. More precisely, the extraction performance will depend on the nature of relations, i. e. how do typical patterns typically appear in free text, when such a relation is stated. Indeed, [15] discuss the nature of binary relations in English, but they use only a small set of relations from a single domain to draw conclusions. Furthermore, they do not consider sentences, where facts with the same relation are encoded several times in one single sentence.

Up to now, we do not use any deeper syntactic features in our system. However, dependent on the nature of relations, it will be helpful to incorporate these types of features as well. Following the idea of [82], we plan to integrate dependency parse trees in our approach. In their work, the syntactic trees are used to segment sentences into several disconnected units before performing NER and RE. In this way, a number of false positives can be filtered out in advance.

After *Text2SemRel* has successfully created a knowledge base, a graph-based and interactive exploration framework is provided to browse through the possibly huge information space. First, *Text2SemRel* computes the knowledge base offline and uses the keywords as attributes for the extracted facts. Other related work [185], in contrast, first applies a standard keyword search to filter out irrelevant documents and performs afterwards the IE step. This strategy might be too restricted, since in this way the set of entities are constrained strictly based on the resulting document set. Our concept-based view is fundamentally different compared to the traditional document-based view. A keyword search in both settings can result in different ER subgraphs, but a quantitative evaluation between the two views is out of scope in this thesis. However, the concept-based view provided by *Text2SemRel* is in line with current visions of a next-generation web such as a web of Linked Data [23] or a Web of Concepts [65].

Chapter 3

Probabilistic Topic Models for Knowledge Discovery in Semantically Annotated Textual Databases

3.1 Overview

Web documents are increasingly annotated with semantic meta data. Semantic annotations play a major role in the realization of the Semantic Web. The annotations are typically from a fixed vocabulary sometimes ordered as a taxonomy or ontology. Web document collections marked with semantic annotations such as classes originating potentially from an ontology, named entities, relations extracted from text, or noisy semantic annotations in form of tags, are expected to represent a major fraction of the web in the future. Following [65], we subsume the just mentioned types of annotations as concepts, semantic annotations or meta data and do not distinguish strictly between them. Semantic annotations describe content concisely and support search and information retrieval. On one side we have high-quality annotations generated by trained professionals. An example here is PubMed[1], a huge biomedical collection of abstracts annotated with Medical Subject Headings (MeSH) terms. On the other extreme are meta data or tags generated by a social network community. Here, tags can be chosen freely, they are of lower quality and contain spelling errors and might have other problems as well.

In this chapter we develop several models that are suitable to model dependencies between the unstructured document representations and their structured annotations [35, 138, 36, 37]. The developed models represent a probabilistic framework that are able to solve a number of important knowledge discovery tasks in semantically annotated document collections. The presented model extensions are based on the framework of Latent Dirichlet Allocation (LDA) [28] or probabilistic topic models, a fully generative model for the generation of document collections. Probabilistic topic models are based upon the idea that documents are generated by a set of topics. A topic is represented formally as a

[1] http://www.ncbi.nlm.nih.gov/PubMed

multinomial probability distribution over terms in a given vocabulary. Informally, a topic constitutes an underlying semantic theme. The notion behind the idea of topic models is that a document usually consists of a large number of words, but might be modeled as deriving from a smaller number of topics. Topic models provide useful descriptive statistics for document collections, which results in applications such as browsing, searching, and assessing document similarity.

We develop two model extensions for the generation of *semantically annotated* document collections: (i) the Topic-Concept model (TC model) [35, 138] (see Section 3.2.3) and (ii) the User-Topic-Concept model (UTC model) [37] (see Section 3.2.4). The first extension models the dependency between documents and their annotations, while the latter one includes an additional user layer. We compare the models on two large document collections to show that they are able to handle different types of semantic annotations: PubMed[2], where annotations originate from a controlled vocabulary and CiteULike[3], where annotations are noisy user-generated tags. PubMed is the largest biomedical document collection today, consisting of about 17 million abstracts most of them annotated with MeSH terms. There are approx. 22.000 MeSH terms arranged in a taxonomy. PubMed annotations are of high quality. On the contrary, CiteULike is a social bookmarking system or collaborative tagging system, where annotations are more noisy. CiteULike allows researchers to manage their scientific reference articles. Researchers upload references they are interested in and assign tags to the reference.

3.1.1 Motivation

The web is undergoing a paradigm shift from a document-centered view towards a more concept-based view. How this concept-based view will exactly look like remains still unclear. Recent advances such as made with Linked Data [23], but also recent position papers from leading search engine providers in which a web of concepts [65] is advocated, give impressions about how the transformation might appear. Semantically annotated documents represent a first important step towards the conversion to a more structured form of the World Wide Web. These type of document collections offer new exciting possibilities for search, retrieval or advanced browsing capabilities.

Besides new possibilities emerging from semantically annotated document collections there are also challenges to face. In collaborative tagging systems, e. g. , semantic annotations are noisy, since users can annotate freely and are not forced to use a specific vocabulary. Annotations in form of tags might be polysemous and different users use slightly different variations of tags to express the same semantics (e. g. consider the tags *information_retrieval*, *information-retrieval* and *IR*). On the other hand, tags must not conform to any common semantic. Also the meaning of a particular tag, such as *to_read*, might be subjective to individuals and does not necessarily express the same shared semantic for the whole community. These aspects make the extraction of meaningful information

[2] http://www.ncbi.nlm.nih.gov/PubMed
[3] http://www.citeulike.org/

from collaborative systems both challenging and rewarding. But also in case where controlled vocabularies are available and annotations are assigned by trained professionals, the situation is far from being perfect [167]. The controlled vocabularies can be very complex and may consist of tens of thousands of unique concepts. As a concrete example consider the MeSH vocabulary with approx. 22.000 different concepts arranged in a taxonomy. Therefore, it is only natural that also human experts annotate inconsistently.

The outlined challenges highlight the urgent need for methods that are able to deal with these inconsistencies and uncertainties. A probabilistic view, as provided by probabilistic topic models, is highly appropriate. Topic models can handle ambiguities in language such as synonymy and polysemy. The presented extensions in this chapter will, in addition to the modeling of documents, also include the modeling of annotations and can thus also deal with ambiguities in annotations. Furthermore, topic models have the compelling property to provide automatically a bird's eye view over a huge document collection by summarizing the content of the collection via occurring topics. In the same way, probabilistic topic models can provide a bird's eye view over a web of concepts by softly grouping related concepts together. As we will see in this chapter, a number of important other knowledge discovery tasks can be solved with the here proposed models.

3.1.2 Related Work

The here presented work is at the intersection of probabilistic topic models and Knowledge Discovery in Textual Databases (KDT) in general. In 1995, [79] laid the foundation for concept-based knowledge discovery in unstructured data. In their work, the authors propose to classify unstructured text documents into conceptual schemes. Afterwards, traditional data mining algorithms executed on the concept level are used for knowledge discovery purposes. While this is also one of the goals in our work, we additionally learn statistical dependencies between the structured annotations and the unstructured textual documents.

Probabilistic Topic Models

In its classical form, LDA [28] is a fully generative model for the generation of discrete data such as document collections. It addresses shortcomings of other dimensionality reduction techniques as discovered in Latent Semantic Indexing (LSI) [67] or Probabilistic Latent Semantic Indexing (PLSI) [101]. LSI computes a Singular Value Decomposition (SVD) to transform the document-term matrix into a lower dimensional space. The rationale behind this dimensionality reduction is that similarities between documents can be estimated more reliably in this reduced latent space. PLSI, the probabilistic extension to LSI, has a statistical foundation but is not truly generative, because there is no natural way to assign probabilities to previously unseen documents [28]. Applications of LDA include automatic topic extraction, query answering, document summarization, and trend analysis. Generative statistical models such as the above mentioned ones, have been proven effective in addressing these problems. In general, the following advantages of topic models are highlighted in the context of document modeling: First, topics can be extracted in a complete unsupervised fashion, requiring no initial labeling of the topics. Second, the resulting representation of topics for a document collection is interpretable. Last but not least, each document is usually expressed by a mixture of topics, thus capturing the topic combinations that arise in documents [101, 28, 89].

The LDA framework can be extended by implying, dependent on the task to solve, another generative process for the generation of documents or other discrete data. [175] introduce the author-topic model that extends LDA by including authorship information. [75] presents a framework to model the generation of text as well as hyperlinks. In the area of social network analysis, [131] represent the author-recipient-topic model, where the modeling of topics is influenced not only by the author of the message but also by the recipient. [47] describe *Nubbi* – Networks Uncovered By Bayesian Inference – a topic model that models entities and their relationships and can thus be used to construct networks.

In what follows, LDA-based approaches are reviewed in the biomedical domain and in the area of social networks, since these areas reflect our current example usages. In the biomedical domain, the classical LDA model has been applied to the task of finding life span related genes from the Caenorhabditis Genetic Center Bibliography [26] and to the task of identifying biological concepts from a protein-related corpus [196]. Another example uses topic models for chemogenomic profiling purposes [81]. In social tagging systems, LDA-based approaches have been barely applied. [154] apply topic models for clustering resources in social tagging systems. Another notable exception is the work of [116], who apply the standard LDA model to tag recommendation. Tag recommendation is about predicting tags, a user assigns to a web resource. The work of [116] was published slightly before our here presented contribution [37]. Another topic model for tag recommendation, published after our here presented contribution, is the tripartite hidden topic model [94]. Similar to [116], the tripartite hidden topic model can only predict tags and more important, they cannot predict tags for a resource for which no tags have been assigned before. While [116, 94] can only apply the models to recommend tags, we present a probabilistic approach that can model every aspect of a collaborative tagging system, i. e. the users, the resources as well as the tags.

Further examples of topic models dealing with meta data in general are [27, 143]. [27] models images and their captions. [143] represent topic models for entities. Another interesting approach is to use existing ontologies to improve the predictive performance of topic models for the words in a document collection [49]. Recently, using topic models for classification purposes has gained a lot of interest. [25, 118] introduce supervised LDA-based approaches for multi-class classification. [153] introduce labeled LDA a model for multi-labeled corpora that incorporates supervision in form of a one-to-one mapping between labels and topics. Note that labeled LDA is published after our presented contributions for muti-label text classification [35, 138].

Collaborative Tagging Systems

Collaborative tagging systems are often seen as a first step towards a more structured web [189] and thus represent another important type of semantically annotated document collections. The Topic-Concept model and User-Topic-Concept model we are going to introduce in this chapter can be naturally applied to collaborative tagging systems, also referred to as folksonomies or social bookmarking systems. Collaborative knowledge platforms have recently emerged as popular frameworks for sharing information between users with common interests. Some popular examples of such systems are Delicious[4], CiteULike[5] or Flickr[6]. A key feature of these systems is that large numbers of users upload certain resources of interest and label them with personalized tags. The resources are in most cases some type of high-dimensional data such as text documents or images. Meaningful annotations adding semantic to the raw resources are given in the form of user specified tags. In contrast to taxonomies, where labels represent ordered predefined categories, no restrictions apply to tags, which are flat and chosen arbitrarily. These free-form strings actually serve the purpose to organize the resources of one single specific user.

The most popular application in these systems is tag recommendation. Some type of collaborative filtering techniques are often applied to this problem [109] or some type of machine learning algorithm such as Support Vector Machines are used for prediction of the most popular tags [98]. Typically, these algorithms are applied on a ”dense” fraction of resources and tags, i. e. the resources and tags have to co-occur a sufficient number of times. [109] present another algorithm for tag prediction, *FolkRank*, which is based on the original PageRank [33] algorithm for ranking web sites. So far, the recommendation algorithms exploit either the provided information from the entire community or the graph structure of the folksonomy to make predictions. On the other hand, in content-based recommendation algorithms, tags are derived from an analysis of the content of the resource. [149] introduce a system, which not only considers the content of a resource, but also takes into account the content of a users‘ desktop to make more personalized predictions.

But tag recommendation is only one of many interesting tasks in these complex systems. Information retrieval issues [102] , the extraction of statistical relations between involved entities in the folksonomy and its mapping to taxonomies [45] as well as knowledge acquisition [169] are also of particular interest. Our contribution provides an integrated view on the just outlined work and applications. Since we define a unified probabilistic model for collaborative tagging systems, we can apply our models in very different scenarios and tasks (see Section 3.3).

3.1.3 Contributions and Outline

Two new probabilistic topic models for the generation of semantically annotated document collections are presented. The approaches model statistic dependencies between unstruc-

[4] http://delicious.com/

[5] http://www.citeulike.org/

[6] http://www.flickr.com/

tured text and their structured counterparts, the annotations. The models can be applied naturally to several different tasks and have direct impact on information retrieval, concept search, prediction of annotations, query expansion, probabilistic browsing, and computing document similarities.

In particular, we present:

1. The Topic-Concept (TC) model, which resembles the generative process of creating a document indexed with semantic annotations. The approach simultaneously models the way how the document is generated as well as the way how the document is subsequently indexed with annotations. The TC model comes with the following features:

   - Topic extraction: In contrast to traditional topic models where topics are modeled by a multinomial topic distribution over words, the Topic-Concept model provides, in addition to the standard topic representation, a multinomial topic distribution over concepts. This results in a soft clustering of concepts that are shared in a common semantic space. A further advantage of the Topic-Concept model is that the representation of topics by words and the representations of topics by concepts are coupled due to the enforced generative process. Thus, for every topic represented by words a corresponding topic represented by concepts is available. The topics represented by concepts give a bird's-eye view of which concepts are semantically related. Furthermore, if the concepts refer to real-world objects, the readability of the topics is greatly improved.
   - Extraction of statistical relationships between involved items: Due to the Bayesian nature of the model, we can easily extract statistical relationships between words, concepts, documents and topics. These properties could be directly applied to query expansion, concept search as well as probabilistic browsing. Note that the relationships are computed based on the shared semantic space and thus go beyond co-occurrence statistics.
   - Semantic interpretation of concepts: By computing for a given concept the most likely topics it is involved in, we get a statistical description about the different contexts in which the concept is discussed.
   - Concept recommendation for new documents: The TC model can predict the most likely concepts for an unannotated document. Another advantage is that the model can include unannotated documents during learning to improve the prediction of concepts. A benchmark with several classification methods on two independent data sets proves our method to be competitive.

2. The User-Topic-Concept (UTC) model extends the former model by including an additional user layer. In addition to the just outlined features of the TC model, this extension comes with the following advantages:

   - Detecting user similarity: similarity between users can be defined elegantly in this model, and one can rank users based on this similarity. The similarity is inferred from the (latent) topic distribution of the users. This can be used e.g. to browse through the libraries of the most similar users.

- Personalized concept prediction: Based on previous citations and concepts from a user, we can predict which new documents he/she would like to cite, and which concepts he/she will use to annotate these documents. By including the learned user-topic distribution into the prediction process we can improve the prediction accuracy in comparison to the TC model.

The rest of the paper is organized as follows: In Section 3.2 we discuss topic models in semantically annotated document collections and introduce LDA and the proposed model extensions. Section 3.3 is devoted to the experimental evaluation of the models. We present results on two different PubMed corpora and on one collaborative tagging system, CiteULike. Our evaluation is based on a quantitative as well as a qualitative comparison. Concerning the qualitative comparison, we present the extraction of the hidden topic-concept structure from these large text collections, the interpretation of concepts via topics, and the extraction of statistical relationships (see Section 3.3.5). The quantitative evaluation is composed of (i) a language model based evaluation as typically done when evaluating topic models, (ii) two challenging multi-label classification tasks (see Section 3.3.3), and (iii) a user-similarity comparison (see Section 3.3.4).

3.2 Topic Models for Semantically Annotated Documents

3.2.1 Terminology

Let $R = \{r_1, r_2, ..., r_{|R|}\}$ be a set of resources or documents, where $|R|$ denotes the number of resources or documents in the corpus. A resource r is represented by a vector of N_r words, where each word w_{ri} in resource r is chosen from a vocabulary of size N. In addition, each resource is annotated with concept labels, where l_{rj} denotes the jth label in resource r. Each resource has M_r such annotations. When a set of users U is given, $\mathbf{u}_r$ denotes the set of users that have assigned resource r. Table 3.1 summarizes the notation used in this chapter.

Notation for Collaborative Tagging Systems

Entities in a social tagging system consist of finite sets of users U, resources R and semantic annotations L in form of tags. Following the notation of [102], a social tagging system or folksonomy F can be represented as a four-tuple:

$$\mathbf{F} = \langle U, R, L, P \rangle, \tag{3.1}$$

where $P \subseteq U \times R \times L$ denotes a ternary relation. Each post p can be represented as a triple:

$$p \subseteq \{\langle u, r, L_{ur} \rangle : u \in U, r \in R, L_{ur} \in L_u\}. \tag{3.2}$$

Count matrices	$C^{TU}, C^{TR}, C^{WT}, C^{LT}$
Dirichlet prior	α
Dirichlet prior	β
Dirichlet prior	γ
Label with index j in resource r	l_{rj}
Labels for resource r	$\mathbf{l}_r$
Number of concept labels	$\|L\|$
Number of labels assigned to a resource r	M_r
Number of resources	$\|R\|$
Number of topics	$\|T\|$
Number of words	$\|W\|$
Number of words in a resource r	N_r
Probabilities of labels given topics	Γ
Probabilities of labels given topic t	Γ_t
Probabilities of topics given resources (or users resp.)	Θ
Probabilities of topics given the resource r	θ_r
Probabilities of topics given the user u	θ_u
Probabilities of words given topics	Φ
Probabilities of words given topic t	ϕ_t
Set of labels in a corpus	L
Set of labels for a user u	L_u
Set of labels for a user u in resource r	L_{ur}
Set of resources in a corpus	R
Set of topic assignments in a corpus	Z
Set of users in a corpus	U
Set of words in a corpus	W
Set of posts in a collaborative tagging system	P
Topic assignment for word i in resource r	z_{ri}
Topic-word assignments in resource r	$\mathbf{z}_r$
Topic assignment for concept j in resource r (drawn uniformly from Z_r)	$\tilde{z}_{rj}$
Topic-concept assignments in resource r	$\tilde{\mathbf{z}}_r$
User assignment to word i in resource r	x_{ri}
User-word assignments for resource r	$\mathbf{x}_r$
Users for resource r	$\mathbf{u}_r$
Word with index i in resource r	w_{ri}
Words in resource r	$\mathbf{w}_r$

Table 3.1: **Symbols associated with the presented models, as used in this chapter.**

Note that $L_{ur} \subseteq L_u \subseteq L$. L_u represents the set of tag labels for a specific user u, while L_{ur} denotes the set of tags assigned by user u to resource r. A tag label l_ur is a specific tag from L_u assigned to resource r.

Algorithm 4: Generative process for the classical LDA model.

```
foreach $topic\ t = 1 \ldots |T|$ do
    sample $\phi_t \sim Dirichlet(\beta)$
end
foreach $resource\ r = 1 \ldots |R|$ do
    sample $\theta_r \sim Dirichlet(\alpha)$
    foreach $word\ w_{ri},\ i = 1 \ldots N_r\ in\ resource\ r$ do
        sample a topic $z_{ri} \sim Mult(\theta_r)$
        sample a word $w_{ri} \sim Mult(\phi_{z_{ri}})$
    end
end
```

3.2.2 Classical Latent Dirichlet Allocation Model

In this section the classical LDA model [28] is introduced and we discuss how LDA can be applied in semantically annotated document collections. LDA, as a fully generative model for the generation of document collections, is based upon the idea that a document or resource is generated by a mixture of topics. Thus, each word in a resource r is drawn from a specific topic. Put in other words, each word w_{ri} in a resource has a specific topic assignment z_{ri}. Each resource is associated with a multinomial distribution over topics. The generation of a resource r is a three step process: First, for each resource, a distribution over topics is sampled from a Dirichlet distribution. Second, for each word in the resource, a single topic is chosen according to the resource-specific topic distribution. Finally, a word is sampled from a multinomial distribution over words specific to the sampled topic. Algorithm 4 depicts this generative process more formally. Θ_r denotes the multinomial distribution for the specific resource r with $\sum_{t=1}^{|T|} \theta_{tr} = 1$. The matrix Θ of size $|T| \times |D|$ stores all topic probabilities given the resources. Similarly, ϕ_t denotes the multinomial distributions over words associated with a topic with $\sum_{w=1}^{|W|} \phi_{wt} = 1$. All probabilities of words given the topics are stored in Φ which is of size $|W| \times |T|$. In Figure 3.1 (Left), the generative process of the LDA model is shown in plate notation. Observed variables are shown in grey. Using the independence assumptions implied by the graphical structure, the joint distribution of a resource r factorizes as follows:

$$p(\mathbf{w}_r, \mathbf{z}_r, \theta_r, \Phi) = p(\Phi|\beta) \prod_{i=1}^{N_r} p(w_{ri}|\phi_{z_{ri}}) p(z_{ri}|\theta_r) p(\theta_r|\alpha). \tag{3.3}$$

Learning the Parameters of the Classical LDA from Textual Collections

In reality, only the resources with their words are observed, thus the task is to extract the underlying topic structure. Exact inference in LDA intractable [28] and the use of approximative inference algorithms is needed. In particular, the task is to infer the word-topic assignments z_{ri} for each word i, the resource-topic distribution θ_r for each resource r

as well as the topic distributions for the corpus ϕ_t for each topic t. The original LDA paper solves this problem by applying mean-field variational methods [28], but other solutions have been proposed as well, such as Gibbs sampling [175] or expectation propagation [68]. Empirical and theoretical comparisons between the various approximation methods used in the context of LDA is a current research issue [140, 9]. In this chapter, we follow the work of [175] and use Gibbs sampling for estimating the desired parameters. Gibbs sampling, a special form of Markov Chain Monte Carlo (MCMC) methods, is an approximate inference method for high-dimensional models. The general idea of MCMC methods is to model a high-dimensional distribution by the stationary behavior of a Markov chain, i. e. samples of the target distribution are drawn based on a previous state, once the method has overcome the ‘burn-in phase’. In Gibbs sampling, each dimension of the target distribution is sampled alternately, conditioned on all other dimensions (see [125] for more details). In LDA, the quantities we are interested in, are the topic assignments z_{ri} for a word w_{ri} in resource r. All other needed variables can be derived from these topic-word assignments. Conditioned on the set of words in a corpus, and the given hyperparameters α, β, we would like to sample the topic assignment z_{ri} for an individual word w_{ri}. The update equation from which the Gibbs sampler draws the hidden variable can be written as follows:

$$\begin{aligned} p(z_{ri}=t|w_{ri}=w,Z_{-ri},W_{-ri},\alpha,\beta) &\propto \frac{p(W,Z|\alpha,\beta)}{p(W_{-ri},Z_{-ri}|\alpha,\beta)} && (3.4)\\ &\propto \frac{C^{WT}_{wt,-ri}+\beta}{\sum_{w'}C^{WT}_{w't,-ri}+|W|\beta}\frac{C^{TR}_{tr,-ri}+\alpha}{\sum_{t'}C^{TR}_{t'r,-ri}+|T|\alpha}, && (3.5)\end{aligned}$$

where the index $-ri$ denotes not to consider the current dimension i in resource r. To derive Equation 3.5 from Equation 3.4 requires to compute the joint distribution $p(W,Z|\alpha,\beta)$, which can be done be marginalizing out Θ and Φ. The interested reader is referred to [96] for a detailed derivation of the here considered quantities. $z_{ri} = t$ represents the assignments of the topic t to the ith word in a resource, $w_{ri} = w$ represents the observation that the ith word is the word w, and Z_{-ri} represents all topic assignments not including the ith word. Furthermore, $C^{WT}_{wt,-ri}$ is the number of times word w is assigned to topic t, not including the current instance w_{ri}. $C^{TR}_{tr,-ri}$ is the number of times topic t has occurred in resource r, not including the current instance. The Dirichlet priors play the role of pseudo counts, assigning non-zero probabilities to topic assignments. α controls the sparsity of the document-specific topic distributions. The larger α is chosen, the more topics will be involved in the generation of a document. In a similar manner, β controls the sparsity of topics, i. e. the larger β is, the more words will become a large probability mass in a specific topic.

Given the estimated topics assignment z_{ri} by the Gibbs sampling procedure, we can now compute the posterior distribution for the multinomial distributions Θ and Φ by using the fact that the Dirichlet is conjugate to the multinomial:

$$\begin{aligned} p(\theta_r|Z,\alpha) &\sim Dir(\theta_r;C^{TR}_{.r}+\alpha) && (3.6)\\ p(\phi_t|Z,\beta) &\sim Dir(\phi_t;C^{WT}_{.t}+\beta), && (3.7)\end{aligned}$$

with $C^{TR}_{.r}$ being the vector of topic counts for the resource r and $C^{WT}_{.t}$ being the vector of word counts for topic t. Using the expectation value for a single variable in a Dirichlet distribution, we get:

$$E[\theta_{tr}] = \frac{C^{TR}_{tr} + \alpha}{\sum_{t'=1}^{|T|} C^{TR}_{t'r} + |T|\alpha} \tag{3.8}$$

$$E[\phi_{wt}] = \frac{C^{WT}_{wt} + \beta}{\sum_{w'=1}^{|W|} C_{w't} + |W|\beta}. \tag{3.9}$$

Gibbs Sampling Procedure Given the update equation 3.5, the Gibbs sampler proceeds as follows: For each document and for each word in a document, the Gibbs sampler starts with a random initialization of topic assignments z_{ri}. The quantities of interest, i. e. the document-topic counts, the sum of document-topic counts, the word-topic counts and the sum of word-topic counts are updated respectively. After the initialization, the Gibbs sampler enters the burn-in phase. In this phase the sampler continues to iterate over all words in all documents, but now draws samples based on $p(z_{ri} = t|w_{ri} = w, Z_{-ri}, W_{-ri}, \alpha, \beta)$. Therefore, the respective counts of the current word under investigation have to be decremented before the sampling step is conducted. After the new topic-assignment for the word under consideration has been assigned, the respective counts are updated. The burn-in phase is repeated until a convergence criterion is reached. In this work we use the log-likelihood of the word observations given the topic assignments and continue until we observe a flattening of the log-likelihood. After the sampler has overcome the burn-in phase, the Markov-chain has reached the target distribution from which we want to draw samples. We define a sampling lag variable that takes every *lag* Gibbs sampling iterations a new sample. We repeat this step a sufficient number of times (ten times in our setting) and average over the samples S. In this way, we ensure that the samples are more decorrelated. Now we can compute the model parameters with help of Equation 3.8 and Equation 3.9.

LDA in Semantically Annotated Document Collections

After having introduced the basic concepts of the LDA model, we discuss how we can apply the classical LDA model in semantically annotated document collections. In general, it is not straightforward to apply this model in these document collections, because LDA assumes that the words of a resource or document originate from one single vocabulary. However, in document collections that have annotations, we basically have two different vocabularies: the first vocabulary originates from the content of the resources R, while the second set originates from the annotations or labels L. There is no way to model the correspondences between the content of a resource and its annotations.

To be able to apply the classical LDA in our given setting we have two options. The first opportunity treats the annotations as observed words and simply merges the two vocabularies. This results in topics that are mixtures of annotations and words, there is no principled way to distinguish between words and annotations. The second option is to treat the annotations as the content of the resources. Thus, a resource r is not represented as a bag-of-words, rather a resource r is modeled as a bag-of-concepts. This ensures that there is no mixture of vocabularies. Indeed, this is done in related work for tag recommendation, see e. g. [116, 94]. However, this approach comes with several drawbacks. First, a number of important discovery tasks such as the extraction of statistical relationships between the content of a resource and its concepts cannot be solved. Second, for recommending concepts for resources, an initial set of concepts has to be given in advance in order to estimate the most likely concepts. If a document has no initial seed of concepts, which will be often the case in real-world applications, there is no way to estimate concepts. Our language model based evaluation in Section 3.3.2 confirms the just mentioned drawbacks. Thus, more advanced models that are able to model semantically annotated documents in a more principled way are necessary.

3.2.3 The Topic-Concept Model

In the last section we have seen that applying the classical LDA to semantically annotated document collections comes with difficulties. Either we use a mixed vocabulary consisting of both words from the resource and semantic annotations or we only use semantic annotations as input and thus do not consider the content of the resources itself. In this section, we present a new topic model, the Topic-Concept model (TC model) [35, 138], which models the content of the resources and their semantic annotations in a principled way. The TC model extends the basic LDA framework by including, in addition to the generation of words, the generation of concepts for the resource. The generation of a resource is first modeled as in the classical LDA model and then the process of indexing the resource with one or several concepts is modeled. The generative process captures the notion that a human indexer or annotator first collects the topics of the resource and afterwards assigns concepts to the resource based on the identified topics. Hereby, the assigned concepts are conditional on the topics that are present in the resource. Note that the concepts can emerge from one single topic of the resource, but also from several topics of the resource. Algorithm 5 summarizes this generative process.

In addition to the three steps needed in LDA for generating a resource, two further steps are introduced to model the process of indexing the resource with semantic annotations. After having modeled the generation of words, an index $i \sim Uniform(1, \dots, N_r)$ is sampled uniformly and the topic assignment z_{ri} of word i is chosen to be the topic assignment $\tilde{z}_{rj}$ for the concept j. Thus, the topic assignment $\tilde{z}_{rj} = z_{ri}$ for the concept j in resource r is based on $\mathbf{z}_r$, the topic assignments of resource r. Finally, each concept label l_{rj} in r is sampled from a multinomial distribution $\Gamma_{\tilde{z}_{rj}}$ over concepts specific to the sampled topic. In addition to the introduced matrices Θ and Φ in the classical LDA model, the matrix Γ of size $|L| \times |T|$ stores the probabilities of concepts given the topics, with Γ_t the multinomial distribution over concepts given a topic t with $\sum_{j=1}^{|L|} \Gamma_{jt} = 1$. The probability distribution

Algorithm 5: Generative process for the Topic-Concept model.

```
foreach topic $t = 1 \ldots |T|$ do
    sample $\phi_t \sim Dirichlet(\beta)$
    sample $\Gamma_t \sim Dirichlet(\gamma)$
end
foreach resource $r = 1 \ldots |R|$ do
    sample $\theta_r \sim Dirichlet(\alpha)$
    foreach word $w_{ri}$, $i = 1 \ldots N_r$ in resource $r$ do
        sample a topic $z_{ri} \sim Mult(\theta_r)$
        sample a word $w_{ri} \sim Mult(\phi_{z_{ri}})$
    end
    foreach label $l_{rj}$, $j = 1 \ldots M_r$ in resource $r$ do
        sample an index $i \sim Uniform(1, \ldots, N_r)$
        set $\tilde{z}_{rj} \leftarrow z_{ri}$
        sample a label $l_{rj} \sim Mult(\Gamma_{\tilde{z}_{rj}})$
    end
end
```

over M_r concepts for the generation of a concept l_{rj} within a resource r is specified as:

$$p(l_{rj}) = \sum_{t=1}^{T} p(l_{rj}|\tilde{z}_{rj} = t)p(\tilde{z}_{rj} = t|\mathbf{z}_r), \tag{3.10}$$

where $\tilde{z}_{rj} = t$ is used as the topic assignment t to the jth concept, and $p(l_{rj}|\tilde{z}_{rj} = t)$ is given by the concept-topic distribution $\Gamma_{\tilde{z}_{rj}}$. It is important to note that by selecting uniformly the topic assignment for a concept out off the assignments of topics in the resource $\mathbf{z}_r$, i. e. $p(\tilde{z}_{rj}|\mathbf{z}_r) = \mathrm{Unif}(z_1, z_2, \ldots, z_{N_r})$, leads to a coupling between both generative components. In this way an analogy between the word-topic representation and the concept-topic representation is created. The principle idea of coupling Θ and Γ has previously been applied successfully to modeling images and their captions [27]. Thus, the generative process of the Topic-Concept model is similar to the Correspondence LDA model proposed in [27] with the difference that the Topic-Concept model imitates the generation of documents and their subsequent annotation, while [27] models the dependency between image regions and captions. As a consequence, in [27] a multivariate Gaussian distribution is used to model image regions, while in the TC model a multinomial distribution is used to sample words given a topic. In Figure 3.1 on the right, the generative process is depicted using plate notation. Observed variables are shown in grey. Using the independence assumptions implied by the graphical structure, the joint distribution of a semantically annotated

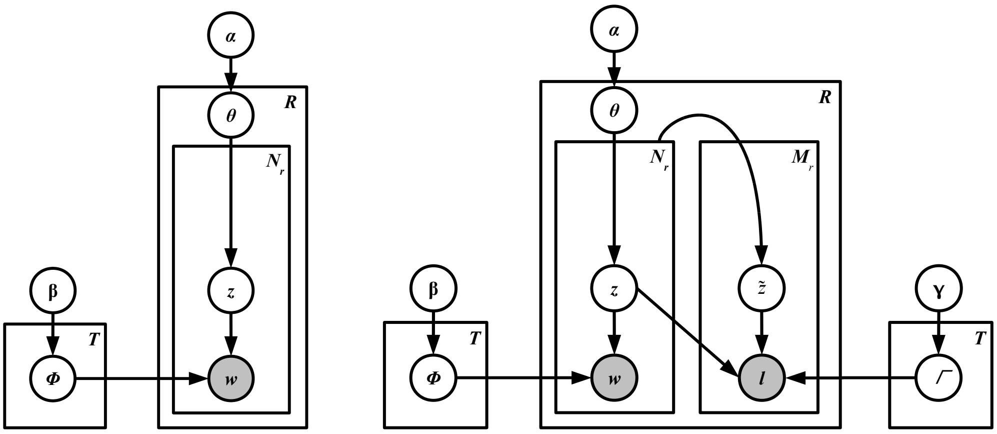


Figure 3.1: **Plate notation of standard LDA and the TC model.** Graphical models in plate notation with observed (gray circles) and latent variables (white circles). **Left:** standard LDA. **Right:** Topic-Concept (TC) model.

resource r factorizes as follows:

$$\begin{aligned} p(\mathbf{w}_r, \mathbf{z}_r, \mathbf{l}_r, \tilde{\mathbf{z}}_r, \theta_r, \Phi, \Gamma) &= p(\Phi|\beta) \prod_{i=1}^{N_r} p(w_{ri}|\phi_{z_{ri}}) p(z_{ri}|\theta_r) p(\theta_r|\alpha) \\ &\cdot p(\Gamma|\gamma) \prod_{j=1}^{M_r} p(l_{rj}|\Gamma_{\tilde{z}_{rj}}) p(\tilde{z}_{rj}|Z_r). \end{aligned} \tag{3.11}$$

With the TC model, we have introduced a probabilistic topic model that models semantically annotated documents in a principled way. The sampling of the topic representation for the concepts is hereby coupled to the word-topic representation, which leads to a corresponding topic representation in both spaces.

Learning the Parameters of the Topic-Concept Model from Text Collections

In the TC model, the quantities we are interested in, are the topic assignments z_{ri} for a word w_{ri} as well as the topic assignments $\tilde{z}_{rj}$ for a concept l_{rj} in resource r. The first part of the generative process is similar to the classical LDA model (see Algorithm 5), therefore the first update equation is similar to Equation 3.5. In the second part, we want to sample topic assignments $\tilde{z}_{rj}$ for a concept l_{rj} conditioned on the set of topic assignments for the labels in the whole corpus, the set of topic assignments made for the words, and the hyperparameter γ. The update equation from which the Gibbs sampler draws the hidden variable yields

$$
\begin{aligned}
p(\tilde{z}_{rj} = t | l_{rj} = l, \tilde{Z}_{-rj}, L_{-rj}, Z, \gamma) &\propto \frac{p(L, \tilde{Z} | Z, \gamma)}{p(L_{-rj}, \tilde{Z}_{-rj} | Z, \gamma)} \\
&\propto \frac{p(L | \tilde{Z}, \gamma)}{p(L_{-rj} | \tilde{Z}_{-rj}, \gamma)} \frac{p(\tilde{Z} | Z)}{p(\tilde{Z}_{-rj} | Z)} \\
&\propto \frac{C^{LT}_{Lt,-rj} + \gamma}{\sum_{l'} C^{WT}_{l't,-rj} + |L|\gamma} \frac{C^{TR}_{tr}}{N_r}. \qquad (3.12)
\end{aligned}
$$

This is obtained by applying the chain rule and the independence assumptions implied by the model. The second term from Equation 3.12 originates from drawing the topic assignment for concept j from $Uniform(1, \ldots, N_r)$. Given the sampled topic assignments $\tilde{z}_{rj}$ for a concept l_{rj}, we can now compute the posterior for Γ by using again the fact that the Dirichlet distribution is conjugate to the multinomial:

$$
p(\Gamma_t | \tilde{Z}, \gamma) \sim Dir(\Gamma_t; C^{LT}_{.t} + \gamma), \qquad (3.13)
$$

with $C^{LT}_{.t}$ be the count vector of labels for the given topic t. The expectation value of the Dirichlet distribution yields:

$$
E[\Gamma_{lt}] = \frac{C^{LT}_{lt} + \gamma}{\sum_{l'=1}^{|L|} C_{l't} + |L|\gamma}. \qquad (3.14)
$$

Together with equations 3.8 and 3.9, we obtain all parameters needed in the TC model.

Gibbs Sampling Procedure As in the Gibbs sampling procedure for LDA, there are three stages: the initialization, the burn-in phase, and the sampling phase. A single Gibbs sampling iteration yields to sample all topic assignments for the words according to Equation 3.5 and to sample all topics assignments for the concepts according to Equation 5. Note that the respective counts of the current word under investigation are first decremented before sampling the new topic assignment for the words. Afterwards, topic assignments for all concepts are sampled with update equation 3.12. Again, the respective counts of the concepts are decremented before sampling the new topic assignment for a specific concept.

3.2.4 The User-Topic-Concept Model

In the last section the TC model was introduced, which handles the generation of concepts based on the topic distribution of a resource r. In this section we introduce another probabilistic topic model that can model another important entity in semantically annotated document collections: users that assign the concepts to the resource. The prime example for this scenario are collaborative tagging systems. Collaborative tagging systems with user generated content have become a fundamental element of websites such as *Delicious*, *Flickr* or *CiteULike*. The here introduced User-Topic-Concept (UTC) model is a

Algorithm 6: Generative process for the User-Topic-Concept model.

```
foreach topic $t = 1 \dots T$ do
    sample $\Phi_t \sim Dirichlet(\beta)$
    sample $\Gamma_t \sim Dirichlet(\gamma)$
end
foreach user $u = 1 \dots U$ do
    sample $\Theta_u \sim Dirichlet(\alpha)$
end
foreach resource $r = 1 \dots R$ and its given users $\mathbf{u}_r$ do
    choose $\Theta_r \sim Dirichlet(\alpha)$
    foreach word $w_{ri}$, $i = 1 \dots N_r$ in resource $r$ do
        Sample a user $x_{ri} \sim Uniform(1, \dots, U_r)$
        sample a topic $z_{ri} \sim Mult(\Theta_{x_{ri}})$
        sample a word $w_{ri} \sim Mult(\Phi_{z_{ri}})$
    end
    foreach label $l_{rj}$, $j = 1 \dots M_r$ in resource $r$ do
        sample an index $i \sim Uniform(1, \dots, N_r)$
        set $\tilde{z}_{rj} \leftarrow z_{ri}$
        sample a label $l_{rj} \sim Mult(\Gamma_{\tilde{z}_{rj}})$
    end
end
```

well-founded probabilistic approach that can model every aspect of a collaborative tagging system. The UTC model captures the notion that several users first cite a resource and afterwards assign concepts to the resource. In the UTC model, the interest of the user is modeled by the assignments of users to words in the resource. The notion behind this assignment is that users might have different objectives to have a resource in their library. Obviously, this is a simplifying modeling assumption. However, this assumption yielded promising results in the past when modeling authors and their interests [175]. Furthermore, as presented in Section 3.3.4, the results for assessing user similarity derived with the UTC model support this modeling choice. Once the UTC model has been trained, the resource-specific topic distribution can be estimated based on a single user. This provides a personalized view on a resource and results in a potential better tag recommendation (see Section 3.3.3). The generative process in the UTC model is formalized by a two-step process where users first cite a resource based on their interests and afterwards assign concepts based on the content of the resource. The way how the assignments of concepts to a resource is modeled, is similar to the TC model. Instead of generating a single model per user, the UTC model is trained globally, reflecting the collaborative aspect of a folksonomy. This enables the UTC model to recommend concepts for a user, that have not been used before by the user to annotate a resource.

Algorithm 6 summarizes the underlying generative process of the UTC model. Each word w_{ri} in a resource r is now associated with two assignments: an user assignment x_{ri} and a topic assignment z_{ri}. We assume that each user u is interested in several topics, thus each user has a multinomial distribution over topics θ_u with $\sum_{t=1}^{|T|} \theta_{tu} = 1$. The matrix Θ of size $|T| \times |U|$ stores the probabilities of topics given the users. First, a user x_{ri} is chosen uniformly for each word of a certain resource. Hereby x_{ri} is chosen from the set of users $\mathbf{u}_r$, the users which cite the resource r. Second, a topic is sampled for each word from the user-specific topic distribution $\Theta_{x_{ri}}$. The difference between the TC model here is that the topics do not originate from a document-specific topic distribution anymore. After the words in a resource r have been generated, the generation of assigned concepts L_r is modeled in the same way as in the TC model. An index $i \sim Uniform(1, \ldots, N_r)$ is sampled uniformly and the topic assignment z_{ri} of word i is chosen to be the topic assignment $\tilde{z}_{rj}$ for the concept j. Finally, each concept label l_{rj} in r is sampled from a multinomial distribution $\Gamma_{\tilde{z}_{rj}}$ over concepts specific to the sampled topic. The difference between the TC model and the UTC model is shown in Figure 3.2 by using plate notation. The joint distribution of a resource r in the UTC model can be represented as follows:

$$\begin{aligned} p(\mathbf{w}_r, \mathbf{z}_r, \mathbf{x}_r, \mathbf{l}_r, \tilde{\mathbf{z}}_r, \Phi, \Gamma, \Theta) &= p(\Phi|\beta)p(\Theta|\alpha) \prod_{i=1}^{N_r} p(w_{ri}|\phi_{z_{ri}})p(z_{ri}|\theta_{x_{ri}})p(x_{ri}|\mathbf{u}_r) \\ &\cdot p(\Gamma|\gamma) \prod_{j=1}^{M_r} p(l_{rj}|\Gamma_{\tilde{z}_{rj}})p(\tilde{z}_{rj}|Z_r). \end{aligned} \tag{3.15}$$

With the UTC model, we have introduced a topic model for semantically annotated document collections that can include user information into the annotation process. The resource-specific topic distribution from the classical LDA or the TC model is replaced by user-specific multinomial distributions and the model captures the notion that a resource gets cited by several users. The generative process of the resource generation is similar to the author-topic model [175] and the generation of the assigned concepts to the resource is similar to the TC model.

Learning the Parameters of the User-Topic-Concept Model from Text Collections

The UTC model contains three continuous random variables, Θ, Φ and Γ. We are interested in the user assignments x_{ri} and the topic-assignments z_{ri} to words w_{ri} as well as the topic-assignments $\tilde{z}_{rj}$ to concepts l_{rj}. Using the independence assumptions implied by the model, we first sample the topic and user assignments to words and afterwards sample the topic assignments to concepts. Conditioned on the hyperparameters α and β, the set of words in the corpus, the set of users, and the set of fixed assignments of users and topics for all other words, we sample the desired assignments using the following Gibbs update for a

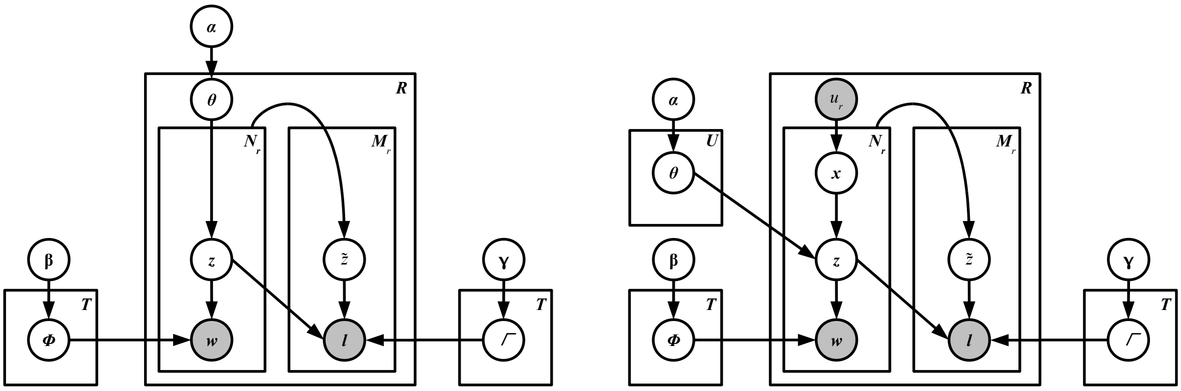


Figure 3.2: **Plate notation for TC and UTC model.** Graphical models in plate notation with observed (gray circles) and latent variables (white circles). **Left:** TC model. **Right:** UTC model.

word w_{ri}:

$$\begin{aligned} p(z_{ri} = t, x_{ri} = u | w_{ri} = w, X_{-ri}, Z_{-ri}, W_{-ri}, U, \alpha, \beta) \;\; &\propto \;\; \frac{p(Z, X, W | U, \alpha, \beta)}{p(Z_{-ri}, X_{-ri}, W_{-ri} | U, \alpha, \beta)} \propto \\ \propto \frac{C^{WT}_{wt,-ri} + \beta}{\sum_{w'} C^{WT}_{w't,-ri} + |W|\beta} \frac{C^{TU}_{tu,-ri} + \alpha}{\sum_{t'} C^{TU}_{t'u,-ri} + |T|\alpha}. \end{aligned} \tag{3.16}$$

The update rule is obtained by applying the chain rule and the independence assumptions implied by the model. A more detailed derivation of the quantity can be found in [175]. The update equation needed for the concept-topic assignments is similar to Equation 3.12, since the generative process of the generation of concepts is similar to the underlying process of concept generation in the TC model. The needed posterior distributions Θ, Φ and Γ are as in the models before Dirichlet distributed, due to the fact that the Dirichlet is conjugate to the multinomial:

$$p(\theta_U | X, \alpha) \sim Dir(\theta_r; C^{TU}_{.u} + \alpha) \tag{3.17}$$

$$p(\phi_t | Z, \beta) \sim Dir(\phi_t; C^{WT}_{.t} + \beta) \tag{3.18}$$

$$p(\Gamma_t | \tilde{Z}, \gamma) \sim Dir(\Gamma_t; C^{LT}_{.t} + \gamma). \tag{3.19}$$

By using the expectation values of the Dirichlet distribution for single variables $E[\theta_{tu}]$, $E[\phi_{wt}]$ and $E[\Gamma_{lt}]$ we obtain all parameters needed in the UTC model.

Gibbs Sampling Procedure The Gibbs sampling in the UTC model works as follows. First, user-word assignments, topic-word assignments and topic-concept assignments, X, Z and $\tilde{Z}$, are initialized randomly. A single Gibbs sampling iteration consists of (i) drawing user and topic assignments for each word w_{ri} sequentially according to update equation 3.16 and (ii) drawing topics assignments for each concept l_{rj} by using update equation 3.12. If the sampler has reached its stationary distribution (i. e. by observing a flattening

	PubMed (genetics-related)	PubMed (random 50K)	CiteULike
Resources	84.076	50.000	18.628
Unique Words	31.684	22.531	14.489
Total Words	4.293.992	2.369.616	1.161.794
Unique Concepts	18.350	17.716	3.411
Total Concepts	912.231	470.101	125.808
Unique Users	—	—	1.393
Total Users	—	—	18.628

Table 3.2: **Corpora statistics used for the evaluation in this thesis.**

of the log-likelihood), we start to draw samples from the posterior distribution. Again, we use the a lag variable *lag* and draw samples after each *lag* iteration for a predefined number of samples S. The samples are averaged over S.

3.3 Experimental Evaluation

In this section we evaluate the proposed topic models and give an impression about the various application areas, the models can be applied to. We start with a description of the experimental setup, where we describe the used corpora and give further training details (see Section 3.3.1). Afterwards we present results for a language model based evaluation as typically done for topic models (Section 3.3.2). In Section 3.3.3 the models are evaluated regarding their predictive power and results for three challenging multi-label classification tasks are presented. Section 3.3.4 presents results for deriving user similarities with the UTC model. The evaluation finishes with the presentation of qualitative results in Section 3.3.5, where learned topic representations and various examples of knowledge discovery applications are presented.

3.3.1 Experimental Setup

Three corpora with a large number of semantic annotations are used for the evaluation of the models. Two corpora originate from the biomedical domain, where resources or documents are annotated with concepts from a terminological ontology. There is no user information available in the data sets, therefore only the TC model is applied to the biomedical corpora. The last corpus is derived from the collaborative tagging system CiteULike[7], where resources are annotated with tags assigned by users. Here we can apply both models.

[7] http://www.citeulike.org/

PubMed Corpora

Two large PubMed corpora previously generated by [141, 142] were used in the experiments. Table 3.2 summarizes the corpus statistics. The first data set is a collection of PubMed abstracts randomly selected from the MEDLINE 2006 baseline database provided by the National Library of Medicine (NLM). Word tokens from title and abstract were stemmed with a standard Porter stemmer [150] and stop words were removed using the PubMed stopword list[8]. Additionally, word stems occurring less than five times in the corpus were filtered out. The collection consists of $|R| = 50.000$ abstracts with a total of 2.369.616 word mentions ($|W| = 22.531$ unique words) and 470.101 concept annotations ($|L| = 17.716$ unique MeSH main headings). We refer to MeSH as a terminological ontology, where relations are partially described as subtype-supertype relations and where the concepts are described by concept labels or synonyms [20]. Note that no filter criterion was defined for the MeSH vocabulary.

The second data set contains $|R| = 84.076$ PubMed abstracts, with a total of 912.231 semantic annotations ($|L| = 18.350$ unique MeSH main headings) and a total of 4.293.992 ($|W| = 31.684$ unique word stems). The same filtering steps were applied as described above. This corpus is composed of genetics-related abstracts from the MEDLINE 2005 baseline corpus. The here introduced bias towards genetics-related abstracts resulted from using NLM's Journal Descriptor Indexing Tool by applying some genetics-related filtering strategies [141]. See [141, 142] for more information about both corpora. In the following, the data sets are referred to as *random 50K* data set and *genetics-related* data set respectively. For the qualitative evaluation the larger genetics-related corpus with all 18.350 unique MeSH main headings was used (see Section 3.3.5).

While the TC model can handle the large number of concepts provided by the MeSH vocabulary, it is difficult to apply the benchmark methods such as a Support Vector Machine or a Naive Bayes classifier to a multi-label classification task of this size. Therefore, we prune each MeSH descriptor to the first level of each taxonomy subbranch resulting in 108 unique MeSH concepts (see Section 3.3.3). In the pruned setting of our task, we have on average 9.6/10.5 (random 50K/genetics-related) pruned MeSH labels per document.

Training Details Parameters for the Topic-Concept model were estimated by averaging samples from ten randomly-seeded runs ($S = 10$), each running over 100 iterations, with an initial burn-in phase of 500 iterations (resulting in a total of 1.500 iterations). We found 500 iterations to be a convenient choice by observing a flattening of the log likelihood. The training time ranged from ten to fifteen hours depending on the size of the data set, the number of used MeSH concepts as well as on the predefined number of topics (run on a standard Linux PC with Opteron Dual Core processor, 2.4 GHz). Instead of estimating the hyperparameters α, β and γ, we fix them to $50/|T|$, 0.001 and $1/C$ respectively in each of the experiments. Hereby, C denotes the size of the vocabulary of the semantic annotations. Therefore, throughout all experiments we use symmetric Dirichlet distributions. The values

[8] http://www.ncbi.nlm.nih.gov/entrez/\query/static/help/pmhelp.html#Stopwords

were chosen according to [175, 89]. We trained the topic models with a predefined number of topics ranging from $T = 200$, $T = 300$, $T = 400$ and $T = 600$ to show that the performance is not very sensitive to this parameter as long as the number of topics is reasonably high. In addition, models with $T = 10$, $T = 50$ and $T = 100$ were trained for the perplexity evaluation in Section 3.3.2.

CiteULike Data Set

CiteULike is a social bookmarking system or collaborative tagging system that allows researchers to manage their scientific reference articles. Researchers upload references they are interested in and assign tags to the reference. Therefore, semantic annotations come in form of noisy tags. CiteULike provides data snapshots of *who posted what* as well as *linkout* data on their web page[9]. The linkout data provides information about the origin of the resource (e. g. a certain article comes from the Science URL). In order to get the content of the resources, i. e. the titles as well as the abstracts, one needs to install several plugins provided by CiteULike. The data snapshot used in our experiments was is from November 13th 2008. We restricted to a reasonable high number of users $|U| = 1393$ and required for the generation of the training set that each resource had to be cited by at least three users. Thus, we wanted to ensure that for the training data set, we obtain a "dense" fraction of resources. Note that we do not use such a restriction for the test set (see next paragraph). In addition, every word token as well as every tag had to occur at least five times in the training data. Word tokens from title and abstract were stemmed with a standard Porter stemmer [150] and stop words were removed using a standard stop word list[10]. Table 3.2 summarizes the corpus statistics. In total our training data set originates from a total of $|P| = 64159$ posts. This comprises $|R| = 18.638$ resources, 1.161.794 words ($|W| = 14.489$ unique words), 125.808 semantic annotations in form of tags ($|L| = 3.411$ unique tags) and 18.628 user mentions ($|U| = 1.393$ unique users). In average each user uses 32 unique tags. The maximum number of unique tag labels for a specific user is 279. The average number of tag assignments per resource for a single user is three. The user id's, resource id's and tags are provided as supplementary data[11].

Test Set for Tag Recommendation We evaluate the here proposed models in a personalized tag recommendation task (see Section 3.3.3). The only restriction for the test set was that a resource had to be posted from a user previously seen in the training set. The same applies to tags. The independent test set consists of 15000 posts.

Training Details Parameters were estimated by averaging samples from ten randomly-seeded runs, each running over 100 iterations, with an initial burn-in phase of 500 for the TC model and 1500 iterations for the UTC model. This results in a total of 1500 and

[9] http://www.citeulike.org/faq/data.adp
[10] http://ir.dcs.gla.ac.uk/resources.html
[11] www.dbs.ifi.lmu.de/~bundschu/UTTmodel_supplementary/info.html

2500 iterations respectively. Again, we found the number of burn-in iterations to be a convenient choice by observing a flattening of the log likelihood. Overcoming the burn-in phase took longer for the UTC model, since a user-topic distribution for each user u has to be estimated as well. Instead of estimating the hyperparameters α, β and γ, we fix them to $50/T$, 0.001 and $1/C$ respectively in each of the experiments (C represents the number of unique tags in the corpus). The values were chosen according to [175, 89]. We trained the topic models with a predefined number of topics ranging from $T = 200$, $T = 300$ and $T = 400$ to show that the performance is not very sensitive to this parameter as long as the number of topics is reasonably high. In addition, models with $T = 10$, $T = 50$ and $T = 100$ were trained for the perplexity evaluation in Section 3.3.2.

3.3.2 Language Model Based Evaluation

We evaluate the TC model and the UTC model in terms of perplexity [12], a quantitative measure for comparing language models that is widely used to compare the predictive performance of topic models (see e. g. [175, 28]). Perplexity, a measurement originating from information theory, measures the ability of a model to generalize to unseen data. Relating to semantically annotated document collections, we use perplexity to measure the ability to predict meta data:

$$Perplexity(L_{test}|R_{train}) = exp\left(-\frac{\sum_{r=1}^{R_{test}} log(p(l_r|R_{train}))}{\sum_{r=1}^{R_{test}} M_r}\right), \tag{3.20}$$

where L_{test} are the concepts in the test set to be predicted and l_r represent the concepts in a certain test resource. R_{train} are the trained parameters, which differ dependent on the used model (see Section 3.2). We compare the classical LDA, the TC model, the UTC model and the so-called Link-LDA model [75] with each other. Link-LDA is a recent extension of the LDA model that not only models text but also hyperlinks between documents. Link-LDA can be used for semantically annotated document collections by treating the semantic annotations as hyperlinks. The generative process of Link-LDA is similar to the generative process of the TC model with the difference, that there is no coupling between the resource-specific topic distribution Θ_r and the concept-specific distribution Γ_r, which means that in Link-LDA concepts can be drawn from latent factors that do not occur in the latent description of the resource.

Perplexity is evaluated on the random 50K data set and CiteULike data set respectively. We partition the data sets into disjoint training (90%) and test sets (10%) and select for each resource in the test set a subset of 50% of the concepts for the evaluation. The remaining 50% of the semantic annotations are used by standard LDA to estimate Γ_r. In contrast, the TC and UTC model exploit the information provided by the resource and first estimate Θ_r, which is estimated online via Gibbs sampling respectively. A small number of iterations $i = 5$ is used to estimate Θ_r. Afterwards the most likely semantic annotations are computed by Γ_r. All perplexity values were computed by averaging over ten different samples. Figure 3.3 plots the perplexity over the held-out semantic annotations under the maximum likelihood estimates of each model for different values of T. Note that a lower perplexity indicates a better annotation quality.

A general trend for both data sets is that using the resources as information clearly adds a benefit (see Figure 3.3). The models including the resource tokens into the computation of the likelihood (Link-LDA, TC model, UTC model) clearly outperform the standard LDA model, which only analyzes the structure in the annotations. On the PubMed corpus the Topic-Concept model performs much better than the Link-LDA model. Recall that the UTC model cannot be applied to this corpus, since no user information is available. The TC model also outperforms Link-LDA on the CiteULike corpus, but the difference in performance is smaller. As T increases, the UTC model gets a better perplexity than the TC model (with a crosspoint at T=100). With T=400 the perplexity of the TC model starts slightly to increase, while for the UTC model the perplexity remains constant.

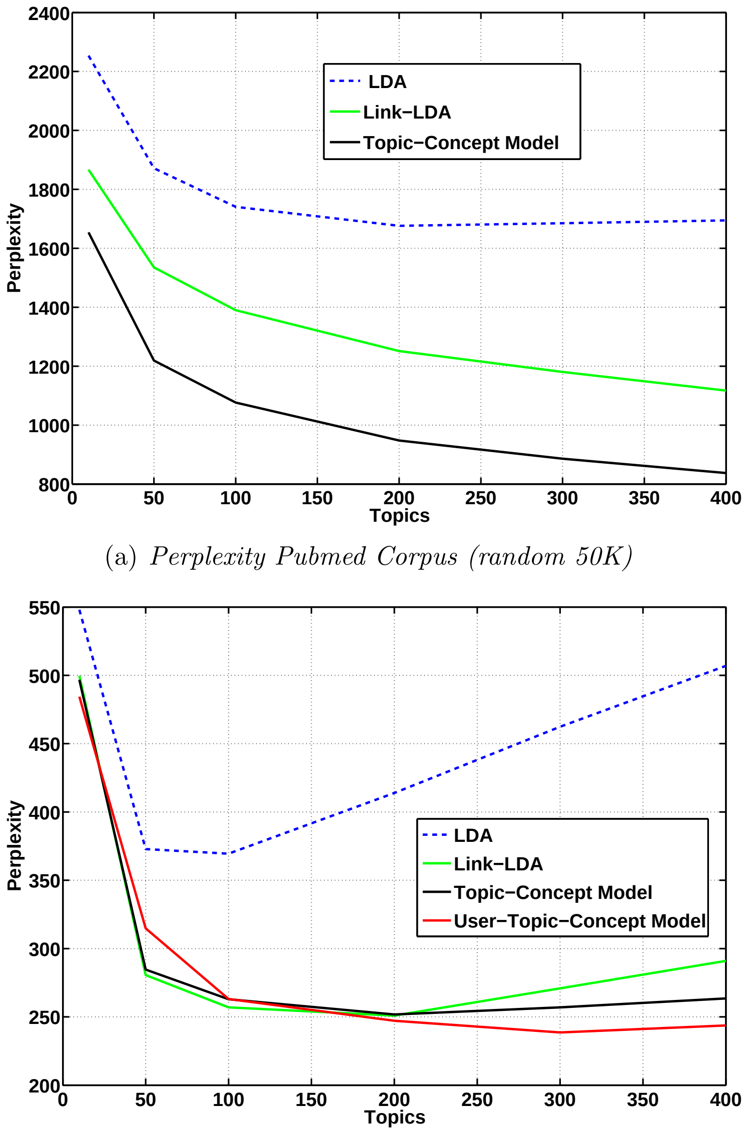
3.3.3 Multi-label Text Classification

To further validate the predictive power of the here presented models, we apply our generative method to three challenging multi-label classification problems and compare the methods with state of the art classification algorithms. First, the TC model is benchmarked on two independent PubMed corpora against (i) a multi-label naive Bayes classifier, (ii) a method currently used by the National Library of Medicine (NLM) and (iii) a state of the art multi-label Support Vector Machine (SVM). The comparison shows encouraging results. After we have proven the predictive power of the TC model against state of the art methods, we compare the TC model with the UTC model on the CiteULike data set in a personalized tag recommendation task.

Results Pubmed Corpora

In this setting, we prune each MeSH descriptor to the first level of each taxonomy-subbranch resulting in 108 unique MeSH concepts. For example, if a document is indexed with *Muscular Disorders, Atrophic [C10.668.550]*, the concept is pruned to *Nervous System Diseases [C10]*. Therefore, the task is to assign at least one of the 108 classes to an unseen PubMed abstract. Note that from a machine learning point of view, this is a challenging 108 multi-label classification problem and corresponds to other state-of-the-art text classification problems such as the Reuters text classification task [121], where the number of classes is approximately the same. In the pruned setting of our task, we have on average 9.6/10.5 (random 50K/genetics-related) pruned MeSH labels per document.

In what follows, we will first describe the used benchmark methods and then present the results for the genetics-related corpus and random 50K corpus. The Topic-Concept model is benchmarked against a method currently used by the NLM [104], which we refer to as *centroid profiling*, a multi-label naive Bayes classifier and a multi-label SVM. For both data sets and all methods, 5-fold cross-validation was conducted.



(a) *Perplexity Pubmed Corpus (random 50K)*

(b) *Perplexity CiteULike Corpus*

Figure 3.3: **Meta data annotation perplexity on test sets.** Note that a lower perplexity indicates a better annotation quality.

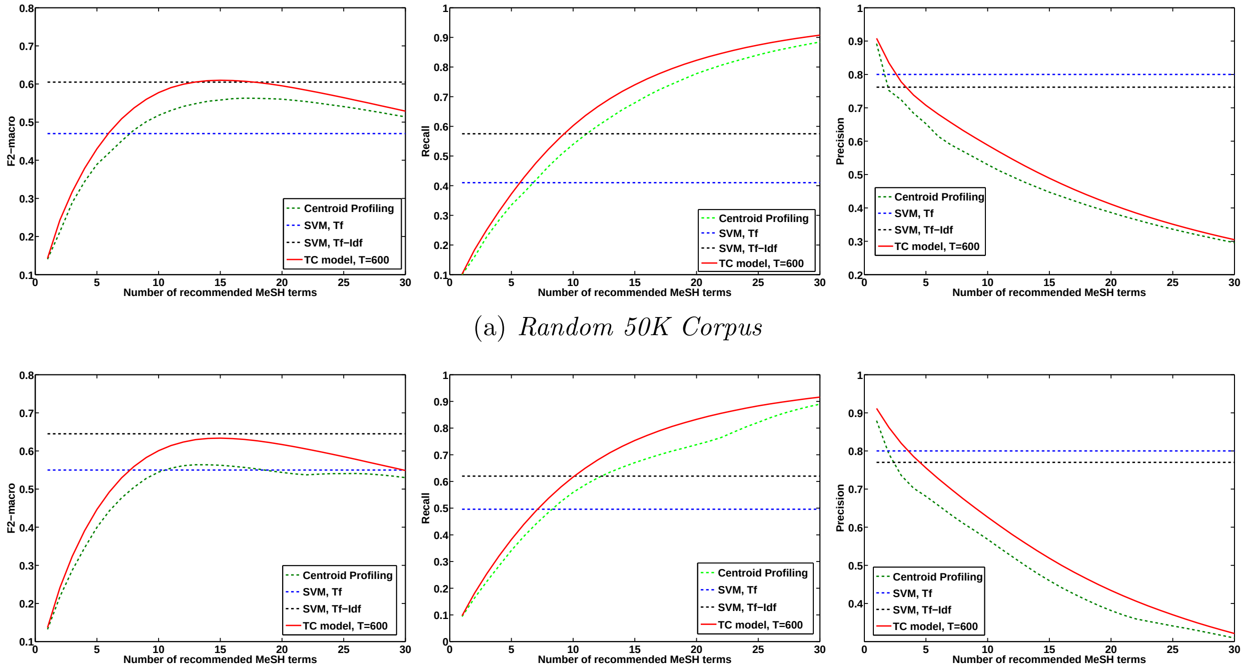
Centroid Profiling In [104] classification is tackled by computing for each word token w_i and each class label l_m, in a training corpus, a term frequency measure $TF_{i,m} = w_{i,l_m} / \sum_{m=1}^{|L|} w_{i,l_m}$ with $|L|$ equals to the total number of concepts. Thus, $TF_{i,m}$ measures the number of times a specific word w_i co-occurs with the class label l_m, normalized by the total number of times the word w_i occurs. As a consequence, each word token in the training can be represented by a profile consisting of the term frequency distribution over all $|L|$ classes. When indexing a new unseen document, the centroid over all profiles for the word tokens in the test resource is computed. This centroid represents the ranking of all class labels for the test resource. This method was chosen, because it is currently used by the NLM in a classification task to predict so-called journal descriptors [104].

Naive Bayes (NB) NB classifiers are a very successful class of algorithms for learning to classify text documents [136]. For the multi-label NB classifier, we assumed a bag of words representation like for the Topic-Concept model and trained it for each of the 108 labels. We used the popular multinomial model for naive Bayes [130].

Multi-Label Support Vector Machine (SVM) The multi-label SVM setting was implemented according to [121]. In this setting, a linear kernel is used and the popular so-called binary method is used to adapt the SVM to a multi-label setting. This setting produced very competitive results on a large-scale text classification task on the RCV1 Reuters corpus [121]. LIBLINEAR, a part of the LIBSVM package [46] is used for the implementation. Two different weighting schemes are evaluated: Term frequency (Tf) as well as cosine-normalized Term frequency-Inverse document frequency (Tf-Idf).

Prediction with the TC Model Section 3.3.1 gives details about the training details. In the TC-model, the prediction of semantic annotations for unseen resources is formulated as follows: Based on the word-topic and concept-topic count matrices learned from an independent data set, the likelihood of a concept label l_{rj} given the test resource r is $p(l_{rj}|r) = \sum_t p(l_{rj}|t)p(t|r)$. The first probability in the sum, $p(l_{rj}|t)$, is given by the learned topic-concept distribution (see Equation 3.14). The mixture of topics for the resource $p(t|r)$ is estimated by drawing for each word token in the test resource a topic based on the learned word-topic distribution $p(w|t)$. The Gibbs sampling is repeated a small number of times ($i = 5$). This online sampling strategy showed good performance for topic model inference on streaming document collections [191]. In contrast to the typical application of topic models in text classification problems, where LDA is usually used for dimensionality reduction purposes [28], the TC model directly predicts a ranked list of concept recommendations.

Evaluation Measure In particular, we are interested in evaluating the classification task in a user-centered or semi-automatic scenario, where we want to recommend a set of concepts for a specific resource (e. g. a human indexer gets recommendations of MeSH terms for a PubMed abstract). In principle, the recommendation task can be considered as a ranking task and well-known information retrieval measures can be used for the evaluation. Indeed, the here proposed topic models return a ranked list of semantic annotations and could be therefore naturally evaluated in a ranking scenario. Unfortunately, it is not straightforward to obtain rankings for classifiers such as SVMs or NB. Thus, we decided to follow the evaluation of [86] and use the well-known F-measure. As done in [86], we average the effectiveness of the classifiers over documents rather than over categories. In addition, we weight recall over precision and use the F2-macro measure, because it reflects that human indexers will accept some inappropriate recommendations as long as the major



(a) *Random 50K Corpus*

(b) *Genetics-related Corpus*

Figure 3.4: **F2-macro, recall and precision plots for the multi-label classification task.** Results are plotted according to the number of top n recommended MeSH terms. In average every document has 9.6/10.5 (random 50K/genetics-related) assignments in our experimental setting. **(a)** Plots for the random 50K data set. **(b)** Plots for the genetics-related data set.

fraction of recommended index terms will be correct [86]. However, in the evaluation on the CiteULike data set (see next Section), the TC model is compared with the UTC model by means of ranking measures.

Results We now discuss experimental results using 5-fold cross-validation. Figure 3.4 plots F2-macro measure, recall and precision against the number of recommended MeSH terms. Figure 3.4(a) shows results for the random 50K data set and Figure 3.4(b) for the genetics-related data set respectively. Our TC model and the centroid profiling method provide as output a ranked list of recommendations. In order to be able to compare these two methods with the other classifiers, a thresholding strategy is needed [190]. We decided to use the simple rank-based thresholding (*Rcut*) [190] and evaluate the results until a cut-off value of 30 (Recall that each document has in average 9.6 (random 50K) and 10.5 (genetics-related) MeSH entries in our experimental setting. The Topic-Concept model was evaluated with two different number of topics on both data sets ($T = 300$, $T = 600$ for the 50K random corpus and $T = 300$, $T = 600$ for the genetics-related corpus). For the sake of clarity, we only show the results for $T = 600$ here, since experimental validation showed that the number of topics is not very sensitive to the overall performance. For the same reason we exclude the NB classifier from the figure (F-measure 0.58 and 0.60

for random 50K and genetics-related). In terms of F2-macro, recall and precision, the Topic-Concept model clearly outperforms the centroid profiling. The naive Bayes classifier already yields quite competitive results. Regarding F2-macro, the TC models reach their optimum at 15 returned recommendations for both data sets (0.61 (random 50K)/0.635 (genetics-related)). At a cut-off value of 15 recommendations, centroid profiling reaches a F2-macro of 0.558 for the random 50K data set (optimum at 17 recommendations with 0.562) and 0.562 for the genetics-related corpus (optimum at 13 recommendations with 0.564). Using a cut-off value which equals to the number of average MeSH assignments (rounded-up) in the two training corpora the F2-macro is for the best TC models 0.59 (random 50K) and 0.61 (genetics-related), while the centroid profiling reaches only 0.517 (random 50K) and 0.55 (genetics-related) at this cut-off value. Note that using the average number of MeSH assignments is the most simple way to determine an appropriate cut-off value. A more analytical way of determining the threshold would be to set up an independent development set for the given corpus and to maximize the F2-macro measure according to the number of recommendations. Other approaches e. g. use a default length of 25 recommended index terms [7] for unpruned MeSH recommendation. The evaluation of the multi-label SVM shows that the performance is very sensitive to the used term weighting scheme (see Figure 3.4). When using Tf-Idf, the SVM is approximately on par with the TC model in terms of F2-macro on both data sets (F2-macro SVM, Tf-Idf is 0.60 (random 50K) and 0.645 (genetics-related)). The SVM is clearly superior in terms of precision due to its discriminative nature. When considering recall, the TC model outperforms the SVM with Tf-Idf, effective from a cut-off value of recommended MeSH terms, which is the average number of MeSH terms in the training corpora.

Semi-supervised Effect of the TC/UTC model Due to their generative nature, the TC/UTC model can naturally include resources that are unlabeled (see also Algorithm 5, 6). The additional resources have a direct effect on estimating Θ and have an indirect effect on Γ, since Γ is coupled to the resource-specific topic distribution. To investigate this effect with respect to classification performance, we treat 95% in the training as unlabeled. Note that the results reported here are based on the random 50K corpus. As a consequence, only 2.000 resources remain semantically annotated (one training split consists of 40.000 resources). The effect on the classification results is as follows (averaged over 5-folds and using the average number of MeSH assignment per resource as cut-off): The TC model trained solely on the 2.000 semantically annotated resources yields a F-2 macro of 0.52, while the results for the TC model trained with 2.000 annotated plus 38.000 unlabeled resources, yields a F2-macro of 0.56. Thus, the unlabeled resources boost the predictive power of the TC model significantly by improving the estimation of the word-topic distribution Θ.

Results Personalized Tag Prediction on CiteULike Data Set

Here, we present results for user-centered tag recommendation and perform evaluation on a post basis. Section 3.2.1 gives an overview for the used terminology. In the last section

NDCG	**@5**	**@10**	**@15**	**all**
Baseline 1	0.04	0.05	0.06	0.19
Baseline 2	0.14	0.20	0.23	0.37
Baseline 3	0.25	0.31	0.34	0.40
TC, T=200	0.29	0.37	0.39	0.47
TC, T=300	0.29	0.37	0.39	0.47
TC, T=400	0.30	0.37	0.40	0.49
UTC, T=200	0.31	0.37	0.40	0.50
UTC, T=300	0.32	0.38	0.41	0.51
UTC, T=400	0.34	0.40	0.43	0.52

Table 3.3: **NDCG evaluation for different number of recommendations.**

we have seen that the TC model is competitive to state of the art classification algorithms. This section is devoted to compare the TC model with the UTC model, when additional user information is available. The evaluation is performed on a post basis, i. e. given an user $u \in U$ and a resource $r \in R$, we want to predict a recommendation or ranking of tags labels $l_{ur} \in L_u$, where L_u denotes the set of tags a user has used so far.

Baselines We follow the baseline methods of previous work on tag prediction [109], but additionally provide personalized versions. The TC and the UTC model are benchmarked against the following standard tag recommendation methods.

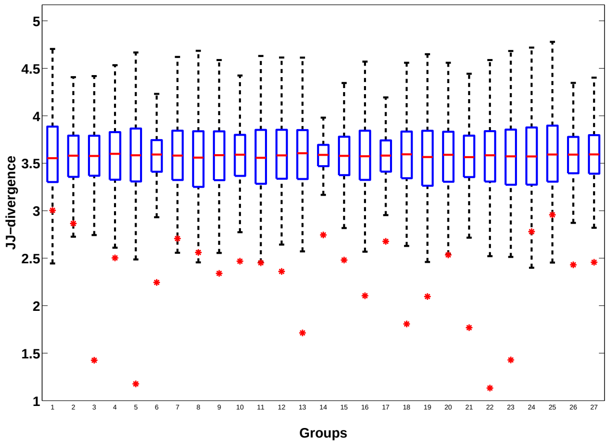
- *Most popular tags:* Tags for a resource r are predicted based on the relative frequency in the collaborative tag system. **(Baseline 1)**
- *Most popular tags with user restriction:* Tags for a resource r are first ranked according to the popularity of tags in the folksonomy and then are reduced to the set of tags L_u. **(Baseline 2)**
- *Most popular tags with respect to the user:* All tags $l_{ur} \in L_u$ are ranked according to the relative frequency as used by u. **(Baseline 3)**

Tag Prediction with the TC and UTC Model For the TC model, the prediction of tags for unseen resources is similar to the prediction for the PubMed corpora: Based on the word-topic and tag-topic count matrices learned from the independent training data set, the likelihood of a tag label $l_{ur} \in L_u$ given the test resource r is $p(l_{ur}|r) = \sum_t p(l_{ur}|t)p(t|r)$. The first probability in the sum, $p(l_{ur}|t)$, is given by the learned topic-tag distribution. The mixture of topics $p(t|r)$ for the resource has to be estimated online. For each resource r, we independently sample topics for a small number of iterations (we used $i = 5$) by using the word counts in Φ from the training corpus.

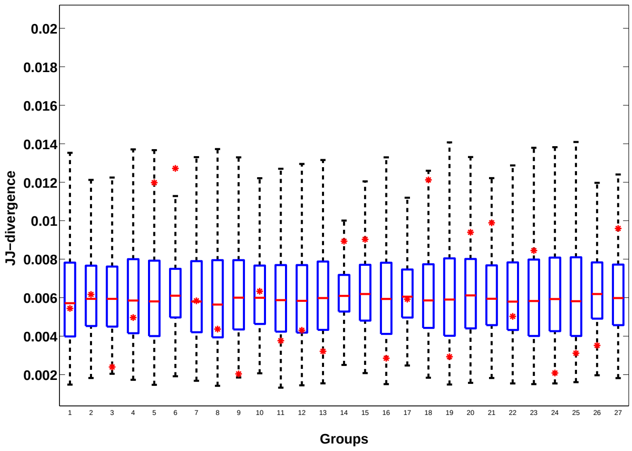
In contrast to the TC model, the UTC model takes into account the user information. The likelihood of a tag label $l_u \in L_u$ in the UTC model is given by $p(l_u|r,u) = \sum_t p(l_u|t)p(t|r,u)$. Again, $p(t|r,u)$ has to be estimated online. Here the mixture of topics for the resource is restricted with respect to the user, i. e. we estimate the topic-distribution for the resource based on the user specific topic-distribution Θ_u. Recall that every post originates from a single user, therefore the estimated topic distribution for the resource under consideration is based on this user. This estimation gives a personalized view on r and thus influences the topic distribution of the resource.

Evaluation Measure In the CiteULike evaluation, we assess the ranking quality of predicted tags, since in this setting each method returns naturally a ranking. We use the normalized discounted cumulative gain (NDCG) [108], a standard measure in the information retrieval community, to evaluate a predicted ranking. The NDCG is calculated by summing over all the 'gains' along the rank list with a log discount factor as $NDCG(\hat{R}) = Z\sum_k(2^{r(k)} - 1)/\log(1+k)$, where $r(k)$ denotes the target label for the k-th ranked item in $\hat{R}$, and Z is chosen such that a perfect ranking obtains value 1. To focus more on the top-ranked items, we also consider the NDCG@n which only counts the top n items in the rank list. These scores are averaged over all posts for comparison. This evaluation scenario reflects that the system provides a list of ranked tags as recommendations for a new cited article and the user chooses tags from the provided list. If the first n ranked tags are the real user assignments for a resource r, reveals an NDCG score of 1. Recall from section 3.3.1 that in average every user has 32 tags.

Results Table 3.3 presents the NDCG scores. The first baseline method performs quite poor, since this model does not take into account which tags a certain user has posted so far. All other methods, i. e. Baseline 2, Baseline 3, the TC and the UTC model take this information into account. The two hierarchical Bayesian models clearly outperform all three baseline methods. Therefore, taking into account the textual resources clearly adds a benefit. The hierarchical Bayesian models are both not very sensitive to the predefined number of topics T, but a slight performance increase can be observed with an increasing number of topics. A major advantage of the UTC model can be observed in the following scenario: 1223 out of the 15000 posts in the test set do only have a resource title and no abstract. Consequently, the number of observed words for making a prediction for the tags drastically reduces. As a consequence it becomes quite difficult to estimate the resource specific topic distribution reliable based solely on the words. Here, the NDCG for the TC model decreases significantly (NDCG all is 0.42 for $T = 200$). The UTC model, in contrast, can make use of the user specific topic distribution to estimate $p(t|r,u)$ more reliably and the NDCG only decreases slightly (NDCG all is 0.47 for $T = 200$).



(a) User profile based on Θ_u from the UTC model



(b) User profile based on the documents of a user

Figure 3.5: **Boxplot for each CiteULike user group in the data set vs. Jeffreys'J divergence.** The red stars indicate the true group divergence. The boxplots for each CiteULike user group are based on 1000 random samplings over all users. In (a) users are represented by the learned Θ_u, while in (b) users are represented by their document vector. The latter method is currently used by CiteULike to assess the nearest neighbors of a user.

3.3.4 Similarity Ranking

The TC and UTC model can be used for a variety of different tasks, a particular interesting one is to assess similarities between items such as words, documents, concepts or users. In this section, our purpose is two-fold: First we show how we can assess similarities with the TC and UTC model in general and second we check if the way how users are modeled in the UTC model is reasonable. As a consequence, but without loss of generality, we concentrate on assessing user similarities in this section. The same approach could be used to assess similarities for semantic annotations or documents, among others.

In order to test if the UTC model is able to identify similar users, we identified all users given in our data set which are members of groups in the CiteULike system. CiteULike groups typically share similar research interests and often belong to one research lab, like for instance the Carnegie Mellon University Human Interaction Institute with a group of 26 users. In our data set there are 488 users out of 1393 which belong to a total of 524 groups (as of November 18, 2008). We excluded all groups with less than five members. This resulted in a total of 27 groups with 160 users. 31 user belong to more than one group and the maximum number of groups for one user is five. First, we derive the similarity between users based on the learned user-topic distributions Θ_u. Since each user profile can be represented as a probability distribution, in particular as a multinomial distribution over topics T, Jeffreys'J-divergence a symmetric version of the Kullback-Leibler (KL) divergence, is used [151]. Jeffreys'J-divergence originates from information theory and is a method to compute the similarity between two probability distributions. To compute the divergence between user u_i and user u_j, we use:

$$\mathbf{Jeffreys}(\theta_{\mathbf{u_i}};\theta_{\mathbf{u_j}}) = \sum_{t=1}^{T} \theta_{tu_i} \log \frac{\theta_{tu_i}}{\theta_{tu_j}} + \sum_{t=1}^{T} \theta_{tu_j} \log \frac{\theta_{tu_j}}{\theta_{tu_i}}, \tag{3.21}$$

while θ_{tu_i} represents the probability of topic t under user i and θ_{tu_j} the probability of topic t under user j respectively. Note that one user can belong to several groups.

Our assumption is that users that share the same group membership should be significantly more similar to each other than users that are randomly chosen and considered as an artificial group. Therefore, we repeated the following procedure for each group: We randomly sampled n users (with n, the size of the group) and computed the mean divergence of this artificial group. This step was repeated 1000 times. Afterwards these results are compared to the true group divergence. Figure 3.5(a) shows the corresponding boxplot for the 1000 samplings for each group. On each box, the central red line is the median, the edges of the box are the 25th and 75th percentiles. The whiskers were chosen such that all data points within $\pm 2.7\sigma$ are considered not as outliers. The stars in the plot indicate the true divergence for each group. All true group divergences fall clearly below the just mentioned percentiles. Furthermore, 20 out of 27 groups are not within $\pm 2.7\sigma$. Figure 3.5(b) shows results, when the documents of a user are used to represent her/his profile. Note that this approach is currently used by CiteULike to assess the nearest neighbors of a user. As can be seen, while still yielding reasonable results for some groups, none of the true group divergences fall out of $\pm 2.7\sigma$. In this setting, for nine user groups the mean of the random sampling is even better than the true group divergence.

3.3.5 Qualitative Evaluation

The purpose of the qualitative evaluation is to highlight the descriptive power of the TC and UTC model for knowledge discovery purposes. The section starts with showing results of the topic extraction from the PubMed and CiteULike corpus. Afterwards, we will show

HIV (Topic 17)				**Phosphorylation (Topic 16)**			
Word	Prob.	Concept	Prob.	Word	Prob.	Concept	Prob.
viru	0.118	Humans	0.06	phosphoryl	0.130	Phosphorylation	0.123
viral	0.064	HIV-1	0.06	kinas	0.118	Prot.-Serine-Threonine Kin.	0.075
infect	0.058	HIV Infections	0.059	activ	0.060	Proto-Oncogene Prot.	0.060
hiv-1	0.047	Virus Replication	0.045	akt	0.060	Proto-Oncogene Prot. c-akt	0.047
virus	0.035	RNA, Viral	0.042	tyrosin	0.036	1-Phosphatidylinositol 3-Kin.	0.047
hiv	0.033	Animals	0.027	protein	0.029	Humans	0.043
replic	0.033	DNA, Viral	0.027	phosphatas	0.025	Signal Transduction	0.038
immunodef.	0.025	Cell-Line	0.023	signal	0.025	Animals	0.028
envelop	0.012	Genome, Viral	0.022	pten	0.024	Protein Kinases	0.021
aids	0.012	Viral Proteins	0.020	pi3k	0.022	Tumor Suppressor Proteins	0.016
particl	0.011	Molecular Sequence Data	0.017	pathwai	0.020	Phosph. Monoester Hydrol.	0.016
capsid	0.011	Anti-HIV Agents	0.016	regul	0.018	Enzyme Activation	0.015
host	0.011	Viral Envelope Proteins	0.013	serin	0.015	Cell Line, Tumor	0.014
infecti	0.010	Drug Resistance, Viral	0.012	inhibit	0.015	Enzyme Activation	0.001
antiretrovir	0.001	Acquired Immunodef. Synd.	0.011	src	0.015	Mice	0.013
Ethics (Topic 6)				**Breast Cancer (Topic 26)**			
Word	Prob.	Concept	Prob.	Word	Prob.	Concept	Prob.
ethic	0.043	Humans	0.150	breast	0.372	Breast Neoplasms	0.319
research	0.039	United States	0.038	cancer	0.323	Humans	0.120
issu	0.023	Informed Consent	0.017	women	0.032	Middle Aged	0.024
public	0.014	Ethics, Medical	0.011	tamoxifen	0.028	Receptors, Estrogen	0.023
medic	0.013	Personal Autonomy	0.001	mcf-7	0.026	Tamoxifen	0.022
health	0.013	Decision Making	0.001	estrogen	0.012	Antineopl. Agents, Hormon.	0.017
moral	0.013	Ethics, Research	0.008	mb-231	0.007	Aged	0.016
consent	0.012	Great Britain	0.008	adjuv	0.007	Carcinoma, Ductal, Breast	0.013
practic	0.012	Human Experimentation	0.007	statu	0.007	Chemotherapy, Adjuvant	0.013
concern	0.011	Public Policy	0.007	hormon	0.007	Mammography	0.012
polici	0.001	Morals	0.007	tam	0.006	Breast	0.012
conflict	0.008	Biomedical Research	0.006	aromatas	0.006	Adult	0.011
right	0.008	Research Subjects	0.006	ductal	0.006	Neoplasm Staging	0.010
articl	0.008	Social Justice	0.006	mammari	0.006	Aromatase Inhibitors	0.009
accept	0.008	Confidentiality	0.006	postmenop.	0.005	Receptors, Progesterone	0.009

Table 3.4: **Selected topics from the TC model with $T = 300$ trained (PubMed corpus, genetics-related).** The fifteen most likely words and MeSH concepts are shown.

examples of the extraction of statistical relationships as well as the semantic interpretation of concepts.

Uncovering the Hidden Topic Structure

Results PubMed Corpora Table 3.4 illustrates several different topics (out of 300) from the genetics-related corpus, obtained from a particular Gibbs sampler run after the 1.500th iteration. Each topic is represented by the fifteen most likely word stems assigned to a specific topic and its corresponding concept representation, here the most likely MeSH main headings. To show the descriptive power of our learned model, we chose four topics describing different aspects of biomedical research. Topic 6 is ethics-related, topic 16 is related to a special biochemical process, namely phosphorylation, and the last two topics represent aspects of specific disease classes. Topic 26 represents a topic centered around breast cancer, while topic 17 refers to HIV. In general, the model includes several other topics related to specific diseases, biochemical processes, organs and other aspects of biomedical research like e. g. Magnetic Resonance Spectroscopy. Recall that the here investigated corpus is biased towards genetics-related topics, thus, some topics can describe quite specific aspects of genetics research. More generic topics in the corpus are related to terms, common to almost all biomedical research areas including terminology, describing experimental setups or methods. In general, the extracted topics are, of course, dependent on the corpus seed. The full list of topics with corresponding word and MeSH distributions is available online as supplementary data[12].

It can be seen that the word stems already provide an intuitive description of specific aspects. However, the resulting word representation of the topics can require some effort to interpret. Clearly, the corresponding concept representation is much more representative and the topics gain more descriptive power by their associated MeSH concepts, providing an accurate description in structured from. Note that the standard topic models are not able to represent corresponding topics of word and concept descriptions (see also Section 3.2.2 for a discussion). In contrast, the TC/UTC model provide a richer representation of topics by additionally linking topics to concepts that can e. g. originate from a terminological ontology. Recall that the coupling of the topics originates from the underlying generative process of the TC/UTC model.

We found that the topics obtained from different Gibbs sampling runs were relatively stable. A variability in terms of ranking of the words and MeSH terms in the topics can be observed, but overall the topics match very closely. For studies about topic stability in aspect models, please refer to [174].

CiteULike Corpus Table 3.5 illustrates four different topics (out of 200) from the CiteULike corpus, obtained from a particular Gibbs sampler run after 2500 iterations. Each table shows the fifteen most likely word stems assigned to a specific topic and its corresponding most likely tags. Again, the coupling between $p(w|t)$ and $p(l|t)$ is a property of the here proposed models and originates from the sampling of a topic for a specific tag based on the topic assignments of the resource (see Section 3.2). To show the descriptive power of our learned model, we chose four topics describing different aspects of the collaborative tagging system. Topic 18 is about the science of networks, while topic 84 reflects a topic about information retrieval. Topic 83 is about social networks and the last illustrated topic 123 about text mining. In general, the extracted topics from CiteULike are quite diverse, ranging from natural language processing specific topics to biomedical topics (e. g. Topic 1, 17 or 127 in the supplementary data[13]). In general, we observe a large fraction of biomedical themes such as systems biology or bioinformatics.

Since the tags in the social bookmarking system CiteULike were chosen freely and by non-professionals, the tags which are ranked highly for a topic were still quite expressive but were somewhat more noisy. An interesting observation can be made: top scored topic tags often contain identical terms with different spellings or terms and their abbreviations

[12] www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/TC_structure.txt

[13] www.dbs.ifi.lmu.de/~bundschu/UTTmodel_supplementary/info.html

Network Science (Topic 18)				**Information Retrieval (Topic 84)**			
Word	Prob.	Concept	Prob.	Word	Prob.	Concept	Prob.
network	0.51	network	0.36	search	0.171	ir	0.21
connect	0.040	networks	0.266	retriev	0.125	search	0.059
complex	0.035	graph	0.009	inform	0.077	information-retrieval	0.054
structur	0.023	complexity	0.009	relev	0.069	retrieval	0.042
topolog	0.020	complex	0.007	ar	0.029	evaluation	0.041
studi	0.019	motifs	0.006	rank	0.026	information	0.036
modul	0.019	social_networks	0.006	feedback	0.024	information_retrieval	0.027
properti	0.017	social-networks	0.005	effect	0.024	relevance	0.023
interact	0.016	modularity	0.005	document	0.023	feedback	0.014
organ	0.016	dynamics	0.005	improv	0.023	ranking	0.011
global	0.013	review	0.005	context	0.021	relevance-feedback	0.011
node	0.013	small-world	0.004	perform	0.018	query	0.011
large-scal	0.012	topology	0.004	techniqu	0.018	personalization	0.009
mani	0.011	systems_biology	0.004	evalu	0.016	user	0.005
robust	0.011	biology	0.004	studi	0.015	semantic	0.005
Social Networks (Topic 83)				**Text Mining (Topic 123)**			
Word	Prob.	Concept	Prob.	Word	Prob.	Concept	Prob.
social	0.155	social	0.21802	text	0.077	nlp	0.14
commun	0.127	community	0.103	extract	0.065	bionlp	0.047
onlin	0.035	privacy	0.029	system	0.034	ner	0.032
particip	0.029	communities	0.025	inform	0.031	text	0.031
share	0.025	communication	0.025	automat	0.030	text-mining	0.020
peopl	0.019	social-networks	0.023	languag	0.0257	information-extraction	0.017
research	0.019	social_networks	0.017	precis	0.025	bib	0.016
media	0.017	online	0.016	task	0.025	ie	0.014
internet	0.017	collaboration	0.014	corpu	0.023	corpus	0.014
discuss	0.016	participation	0.014	entiti	0.023	textmining	0.014
member	0.013	hci	0.009	natur	0.022	text_mining	0.011
relationship	0.013	internet	0.008	evalu	0.021	new	0.009
technolog	0.012	tagging	0.006	sentenc	0.019	extraction	0.009
contribut	0.012	blog	0.006	process	0.019	umls	0.008
group	0.012	blogging	0.006	word	0.018	annotation	0.008

Table 3.5: **Selected topic from a UTC model with $T = 300$ (CiteULike corpus).** For each topic the five most probably words and tags are listed

(consider e. g. IR vs. information retrieval). On the one hand, this reflects the nature of collaborative tagging systems, where different users use different variations to express the same semantics, but on the other hand, it also shows that the models can capture the meaning of the annotations by clustering these terms softly in the same semantic space.

Note that in the UTC model, we can also represent the most likely users given the topics $p(u|t)$. This information gives us an overview about which users are mainly interested in which topics. Information about $p(u|t)$ in CiteULike provides interesting insights about the main research interests of users. The most likely users given the topics for a UTC model with $T = 200$ are also provided as supplementary data.

Extraction of statistical relationships

Besides uncovering the hidden topic-concept structure, we can apply the model to derive statistical relations between items involved in the generative process. One such example is

Diseases	
Myelodysplastic Syndromes (208)	Pulmonary Embolism (39)
acut aml bcr-abl blast chronic cml flt3 hematolog imatinib leukaemia leukem leukemia lymphoblast marrow mds myelodysplast myeloid patient relaps syndrom	activ associ case clinic diagnos diagnosi diagnost factor incid men mortal patient platelet preval protein rate risk studi women year
Drugs & Chemicals	
Erythropoietin (85)	Paclitaxel (309)
abnorm anaemia anemia caus cell defect defici disord epo erythrocyt erythroid erythropoietin g6pd hemoglobin increas model normal patient sever studi	advanc agent anticanc cancer chemotherapi cisplatin combin cytotox drug effect median paclitaxel patient phase regimen respons sensit surviv toxic treatment

Table 3.6: **Selected MeSH concepts and their most likely words**. Selected MeSH concepts from the *Disease* and the *Drug & Chemicals* subbranch with the 20 most probable word stems estimated based on a topic-concept model learned from the genetics-related corpus ($T = 300$). The font size of each word stem encodes its probability given the corresponding MeSH concept. The number in brackets is equal to the number of times, the MeSH terms occurs in the corpus

the extraction of relationships between concepts and words, thus bridging the gap between natural free text and the structured semantic annotation. The derived relations could be e. g. used for improving word sense disambiguation [105]. In Table 3.6, four MeSH concepts from the *Disease* and the *Drug & Chemicals* subbranch and their twenty most probable word stems are shown. For each MeSH concept, the distribution over words is graphically represented by varying the font size for each word stem with respect to the probability. Given a concept c, the conditional probability for each word is estimated by $p(w|c) \propto \sum_t p(w|t)p(t|c)$, which is computed from the learned model parameters. The word distributions describe the corresponding MeSH concept in an intuitive way, capturing the topical diversity of certain MeSH concepts.

Note that there are many other opportunities to access statistical relations between MeSH concepts and words. One could e. g. use measurements like co-occurrence or χ^2 statistics. It may be that the TC/UTC model captures relationships that can't be captured in a simpler way, but this evaluation is out of scope of the here presented work. We provide all word clouds for all MeSH terms occurring in the corpus from the *Disease* and the *Drug & Chemicals* subbranch as supplementary data[14].

[14] www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/

It is important to note that we highlight only one particular example of statistical relationship extraction here. Another interesting use case that can be solved with the here introduced model is given a set of concepts, what are the most likely resources? In this way, resources or documents can be retrieved that are semantically related with the given concepts, but the resources do not necessarily have to be annotated with the concepts. Recall that even trained professionals annotate inconsistently and therefore such a probabilistic view is highly suitable.

Semantic Interpretation of Concepts

Another important use case is the semantic interpretation of concepts. Given a concept we might not know anything about the concept, e. g. what it means and in which contexts it is discussed. The extraction of statistical relationships as done in the previous section already gives an initial description of a concept (see Table 3.6). But with the TC/UTC model we can additionally access the most likely topics given a concept $p(t|l)$ and can thus get an idea about the different contexts in which the concepts plays a role.

As a concrete example, we estimate the most likely topics given a specific MeSH concept with respect to a seed corpus. This results in a fast overview over the topics in which a specific MeSH term is most likely to be involved in. Table 3.7 shows two such examples extracted from the genetics-related corpus. Here, the topics are presented by the most likely word stems, but the corresponding concept representation can be used as well (see supplementary data[15], concept representation of the topics). The first example shows the three most likely topics for the MeSH term *myelodysplastic syndromes*. Myelodysplastic syndromes, also called pre-leukemia or ‘smoldering’ leukemia, are diseases in which the bone marrow does not function normally and not enough blood cells are produced [145]. This fact is reflected by the most likely topic for this MeSH term (Table 3.7, Topic 46). Furthermore, a state of the art treatment of this disease, is bone marrow transplantation. First, all of the bone marrow in the body is going to be destroyed by high-doses of chemotherapy and/or radiation therapy. Then healthy marrow is taken from a donor (i. e. another person) and is given to the patient [145]. This is described by the second most likely topic (Table 3.7, Topic 75). Topic 25 constitutes that Myelodysplastic syndromes have a genetic origin and that gene and chromosome aberrations are a likely cause of this disease [145].

The second MeSH term in table 3.7, *Erythropoietin* (EPO), is a hormone which is produced by the kidney and liver. It is known to regulate red blood cell production. In the mined genetics-related corpus, the most likely topic (Table 3.7, Topic 177) states that erythropoietin could be used as a treatment during malaria infection [44] and this is indeed a current issue of ongoing research [21, 187]. Erythropoietin is known to directly promote the generation of neuronal stem cells from progenitors, which is reflected by Topic 14. Last but not least, Topic 140 provides information about the gene regulatory context of EPO. NF-kappaB, e. g. , regulates EPO [43], while EPO in turn regulates expression of c-jun and AP-1 [170].

[15] www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/TC_structure.txt

MeSH term	Topic	Word stems
Myelodysplastic Syndromes	Topic 46 ($p = 0.20$)	leukemia acut myeloid aml mds lymphoblast leukaemia blast leukem patient myelodysplast marrow syndrom malign flt3 bone promyelocyt hematolog mll granulocyt
	Topic 75 ($p = 0.02$)	transplant donor recipi graft stem allogen reject autolog cell immunosuppress allograft marrow surviv hematopoiet condit receiv acut gvhd engraft diseas
	Topic 25 ($p = 0.01$)	chromosom aberr transloc cytogenet delet abnorm rearrang genom karyotyp gain loss region arm breakpoint trisomi mosaic duplic cgh case imbal
Erythropoietin	Topic 177 ($p = 0.30$)	defici adren anemia malaria parasit plasmodium mosquito falciparum erythrocyt cortisol erythropoietin caus g6pd insuffici adrenocort acth anaemia epo anophel develop
	Topic 14 ($p = 0.14$)	cell stem progenitor hematopoiet differenti embryon lineag hsc adult marrow bone erythroid cd34+ precursor potenti cd34 marker hematopoiesi msc self-renew
	Topic 140 ($p = 0.07$)	activ nf-kappab factor nuclear transcript express cell induc inhibit constitut ap-1 regul c-jun suppress p65 kappa curcumin transloc nfkappab c-fo

Table 3.7: **Selected MeSH concepts with most likely topics.** Concepts are from the *Disease* and the *Drug & Chemicals* subbranch with the three most probable topics estimated based on a topic-concept model learned from the genetics-related corpus ($T = 300$). Topics are illustrated here by the twenty most probable word stems.

A full list of all MeSH terms occurring in the corpus and its most likely associated topics is available online[16].

3.4 Conclusion

Summary This chapter presented probabilistic topic models for the generation of semantically annotated document collections. We introduced two novel topic models, the Topic-Concept (TC) and the User-Topic-Concept (UTC) model. The Topic-Concept model learns relationships between words, concepts, documents and topics from large annotated text corpora. The User-Topic-Concept model presents an extension of the former model by including an additional user layer into the generative process of a semantically annotated resource. With the UTC model we are able to exploit the complete spectrum of information available in collaborative tagging systems. These systems represent important sources

[16] www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/mesh_associated_topics.pdf

for populating semantic web applications [45].

The underlying idea of both models is to map all involved entities, i. e. the users (if available), the resources, the words and the assigned semantic annotations into a common lower dimensional latent topic space. In this way the great variety of ambiguous information inherent in semantically annotated document collections can be drastically reduced. The here proposed models can be applied naturally to various tasks.

As a quantitative result we showed that the here proposed models provide a better meta data annotation quality in terms of perplexity compared to the standard LDA framework. The results in three challenging multi-label classification tasks are encouraging and show that the TC/UTC model can compete with state of the art classification algorithms. Furthermore, the models can naturally include unannotated resources due to their generative nature and thus are suitable for semi-supervised settings. In the similarity ranking evaluation, the user similarity derived from the User-Topic-Tag LDA fits well with the structure of user groups in CiteULike.

A number of further important knowledge discovery tasks can be solved with the TC/UTC model. Among others, the topic representation is richer compared to the standard LDA framework. If the concepts refer to real-world objects, as is the case for the PubMed corpora, the readability and interpretability of topics is greatly improved. Last but not least, descriptive interpretations of concepts can be obtained with the here presented model extensions.

Discussion The topic models presented in this chapter try to imitate the process of indexing a resource with concepts. While we tried to set up the generative processes analogue to a 'real' indexing process, several simplifying modeling assumptions have been made. For instance, even though yielding convincing results, one such assumption is that the interest of users can be expressed by the assignment of users to single words. An issue of ongoing research will be to investigate more ways to resemble the generation of semantically annotated document collections.

Regarding the multi-label classification task, the resource-specific topic distribution θ_r for an unseen test document was sampled by using a fixed, small number of Gibbs iterations ($i = 5$). The effect of the number of Gibbs sampling iterations with respect to the classification performance is an issue of future research. More importantly, θ_r was estimated independently per document. We decided to follow this strategy, since considered in the context of an online recommendation system, it represents the most realistic scenario. But in principle, the resource-specific topic distribution can be estimated also jointly for a whole test set of documents. This might have a beneficial effect on estimating θ_r and in turn may also increase the recommendation performance.

Considered from an application perspective, there are several future directions worth exploring with the presented topic models. For instance, based on the encouraging results in assessing user similarities, a particular interesting application that would be important to analyze quantitatively, is to derive the relatedness of concepts with the TC/UTC model. Studying the relatedness of concepts (including tags in social bookmarking systems) is an active area of research and it will be very interesting to compare similarities obtained by the TC/UTC model with existing measures. E. g. , [45] provide a thorough study of tag relatedness in a social bookmarking system. Also recently, explicit semantic analysis (ESA) [84] has been proposed as a way to compare documents by translating text into a weighted vector of concepts by using large knowledge bases. One of the advantages highlighted in the context of ESA is the use of explicit human designed knowledge. Similar to ESA, the TC/UTC model exploits human-defined concepts and can rank these given a textual fragment. This makes a future comparison of the two methods worth exploring. Another interesting future direction is to apply the TC/UTC model in a probabilistic information retrieval setting. The TC/UTC model can easily handle queries that are loose mixtures of concept and words. The mixture of words and concepts can be translated directly into the latent topic space and the most likely documents can be retrieved. Inconsistencies in the semantic annotations can be compensated by assessing the similarity of the query and the document in the lower dimensional semantic space.

Chapter 4

Digging for Knowledge with Information Extraction: A Case Study on Human Gene-Disease Associations

4.1 Overview

As of December 2009, there are 5414 journals indexed in the world's largest biomedical database[1] PubMed. From the year 2000 to 2008 the number of articles stored in PubMed almost doubled. With each new published article, a cohort of new facts is introduced to the public. This immense growth of literature and experimental data in the biomedical domain calls for automatic methods that extract these myriads of potential new findings.

In the last two chapters statistical methods have been proposed that are able to extract information efficiently from textual data sources. In this chapter *Text2SemRel* (see Chapter 2), our proposed fact extraction system, is applied to extract semantic or typed relations between genes and diseases from a large textual repository. The semantic relations are the same used in the experimental evaluation of *Text2SemRel* in Chapter 2, Section 2.4.3. Thus, the relations are classified into several biomolecular conditions, describing a wide variety of molecular conditions. They range from genetic to transcriptional and phosphorylation events. The resulting knowledge base contains far more potential disease genes than currently stored in curated databases. This **L**iterature-derived **H**uman-**G**ene-**D**isease **N**etwork (LHGDN) is subject of further analysis in this chapter.

The LHGDN is compared against several curated state of the art databases. Section 4.2 introduces the chosen curated databases. Afterwards, the LHGDN is analyzed with regard to properties and characteristics of known disease genes. The experimental analysis shows that the facts extracted from literature are of high quality. Furthermore, a careful statistical analysis gives interesting insights about how facts are published. The results of

[1] http://www.nlm.nih.gov/bsd/num_titles.html

the analysis are discussed in Section 4.4.

4.1.1 Motivation

A particular important piece of biomedical knowledge is about how genes associate to diseases. Our knowledge about the possible causes of complex diseases is still limited and for many diseases there are still no proper cures. The visibility of facts known to individual researchers, appears to be very restricted compared to the whole accessible knowledge. This phenomenon, termed as knowledge pockets [62], promotes the limited knowledge about complex diseases. Even though several gene-disease association repositories do exist on the web, they all have a special focus and thus are highly diverse and non-redundant (see Section 4.2). More importantly, the databases use different controlled vocabularies that are suitable for their purposes. As a consequence, data integration issues are another major drawback. However, many diseases are caused by the effect of several genes and thus a unified view on human gene-disease associations will help to improve the understanding of complex diseases tremendously.

Information Extraction techniques have the great potential to help to overcome the problems just outlined. Given that all or at least most of the available knowledge is somewhere written in textual form and thus can, in principle be extracted, text mining approaches have a compelling property: they can alleviate the data integration problem by defining consistent controlled vocabularies in advance and search subsequently through all available unstructured data sources.

Haunting for disease genes Starting in the 80s with almost no knowledge about gene-disease associations, the number of discovered disease genes is steadily increasing [112]. Similar to the intense discussion in the 90s about the true number of genes for the human organism, it is not clear how many disease genes there are. Of course, this number will be biased heavily by the definition of a disease gene. However, the results presented in this chapter indicate that the true number of disease genes might be much larger than the number shown in [112]. One reason for this difference is that human curators have to cope with tons of new biomedical findings. As a consequence, curated repositories often have a noticeable delay in updating and are far from being complete [146]. We refer to disease genes as genes that are either involved in the causation of or associated with a disease [112]. Traditionally, knowledge about genes that cause or predispose to a disease have been the focus of research. But genes must not necessarily cause or predispose to a disease in order to be of interest. Recently, researchers became aware that it is also of major importance to study genes that are simply associated with a disease. E. g. , knowing genetic variations of genes associated with diseases or knowing that the protein product of the gene is increased or decreased in a disease is crucial for the development of new molecular markers. Those have the great potential to personalize medicine. For instance, while prognostic markers are used to predict survival rates, predictive markers are helpful to predict response to drug treatments. Not surprisingly, this research direction is also a major focus of the pharmaceutical industry. See [4] for a recent review and examples, where biomarkers are implemented successfully.

4.1.2 Related Work

In the broader sense, the here presented work is related to the field of translational bioinformatics [112]. In the narrower sense, the here presented work is related to systems that mine the biomedical literature in a large-scale manner, which is the focus of this section. A detailed overview of information extraction approaches in general is given in Chapter 2, Section 2.1.4. Recent reviews about advances in biomedical text mining can be found in [166, 85].

Large-scale Biomedical Information Extraction [160] analyzed their own professional library system at Biogen Idec., giving insights into information needs of researchers in drug development. They show that drugs, diseases and genes are the top ranked entities, industrial researchers are asking for in this area. Concerning the type of sources, journal articles are the top resources, followed by competitive intelligence and patent sources. In the same spirit, the well-known GeneWays system [165] mines the literature for relations between drugs, genes and diseases, while another system, Chilibot, extracts protein/gene interactions from text and considers the context in terms of keywords [50]. Approaches that mine gene-disease associations from text are most related to our work. Following our contribution [34], [146] use the GeneRIF database to extract relations. However, they do not use typed relations and do not filter for negative associations. [55] suggest the need for smart filtering functions for reducing false positives with the help of machine learning, but also do not use typed-relations. In contrast, [158] classifies relations into whether a gene causes, predisposes or is simply associated with a disease. Recall that the relations defined in our contribution focus on the molecular level. A classification scheme that is more similar to the one used in our work is introduced in [54]. The relations are classified into six different types: study description, genetic variation, gene expression, epigenetics, pharmacology, and clinical marker . However, the proposed system is only able to extract relations between genes and one specific disease, namely prostate cancer. The BITOLA system uses biological background knowledge about chromosomal locations to rank candidate gene-disease relations extracted from the literature [103]. Finally, we would like to highlight PepBank, a large database of peptides that was obtained with information extraction techniques [172].

Work comparing curated knowledge sources with knowledge bases originating from text mining is found rarely in the biomedical domain. A recent example extracts whole interactomes (interactions between different molecules such as proteins, lipids and nucleic acids) from PubMed for the human and mouse organism [106]. Furthermore, an excellent work compares curated and text mining sources in the protein-protein domain [82].

Also related to the here presented contribution is research that uses information extraction techniques to extract networks of interacting entities in order to perform statistical analyzes afterwards. [122] first extract gene-disease relations via statistical co-occurrence analysis over 2.6 million PubMed abstracts, what resulted in a total of 4195 genes linked to 1028 diseases. Second, pathway annotations of the genes are used to annotate diseases with pathways. Afterwards a disease network is build. This is done by linking two diseases if they share at least one common pathway. This disease network is subject of further network analysis. [147] first extract a gene interaction network from the literature in order to identify new potential disease genes with the help of network measures such as centrality. [78] build a literature-derived interactome and combine this network with a network obtained from Yeast2Hybrid studies. The union of these networks is used to deviate properties of disease genes. [62] use extracted relations to study the behavior of the growth of knowledge and thus is close to our contribution in Section 4.4.3.

4.1.3 Contributions and Outline

The main contributions that will be presented in this chapter are the following:

- *Text2SemRel* is used for constructing a knowledge base consisting of facts centered around gene-disease relations from a large textual repository. The resulting knowledge base is publicly available[2] [83]. In addition, the LHGDN is integral part of the Linked Life Data[3] project.

- A comparison against several state of the art gene-disease association databases is provided.

- The properties of the LHGDN are analyzed in depth.

The rest of the chapter is organized as follows: In Section 4.2, current state of the art gene-disease databases are reviewed. Section 4.3 gives details about the creation of the LHGDN. Finally, Section 4.4 presents the results of the analysis of the LHGDN.

4.2 State of the Art Curated Human Gene-Disease Association Databases

In this section, we will briefly review the chosen databases that are compared with the LHGDN. All databases, except the LHGDN, are curated by humans. At the end of this section, a new database (named ALL in this chapter) is introduced, which is an attempt to unify the single data sources to provide a more complete and comprehensive view of human gene-disease associations [83]. The LHGDN, introduced in this chapter, is part of this database. To the best of our knowledge, ALL represents the most complete view of human-gene disease associations available today [83]. The basis of this new data source is formed by a gene-disease association ontology created by two biomedical experts (Authors of the paper [83]: Furlong, L. and Bauer-Mehren, A.) that unifies the different focuses of

[2] http://www.dbs.ifi.lmu.de/~bundschu/LHGDN.html
[3] http://linkedlifedata.com/sources

the various databases (see Figure A.1). However, a detailed discussion of this new data source is out of the scope in this thesis.

The databases were chosen according to following criteria:

- Different scopes of the databases. To understand complex diseases such as cancer, it is vital to have a complete view of all known gene-disease associations that unify different scopes. E. g. while one database focuses on Mendelian diseases, another database stores gene-disease associations with respect to environmental factors.
- The databases must use a controlled vocabulary that can be mapped to the UMLS[4] (Unified Medical Language System) to improve interoperability among the different sources. In addition, we tried to use databases that use the same vocabulary for disease and gene entities as far as possible.
- Furthermore, databases were chosen according to characteristics such as reliability, acceptance in the biomedical community and public availability.

OMIM® - Online Mendelian Inheritance in Man [92] The OMIM database focuses on inherited or heritable diseases and is seen as one of the most comprehensive data sources about gene-disease associations. However, only recently OMIM started to include diseases with complex traits and is far from being complete [146]. Gene-disease associations were obtained by filtering for associations of type 'phenotype' (data[5] downloaded on June, 6th 2009). In total, 2.198 distinct genes and 2.473 distinct disease terms, comprising 3.432 gene disease associations, were obtained. After merging disease vocabularies (for more details see [83]), the OMIM net contained 2417 distinct diseases. The resulting associations were classified as *Marker* in the gene-disease association ontology [83].

UniProt - Universal Protein Resource [5] UniProt is a database containing curated information about protein sequence, structure and function. Moreover, it provides information about the functional effect of sequence variants and their associations to disease phenotypes. This information from UniProt release 57.0 (March 2009) was extracted as described in [16]. All protein identifiers were converted to Entrez Gene identifiers in order to allow integration with the other data sources. This mapping is straightforward and only some terms cannot be mapped. Using this approach, UniProt provided 1.746 distinct gene-disease associations for 1.240 distinct genes and 1.475 distinct diseases. All gene-disease associations were classified as *GeneticVariation* in the gene-disease association ontology [83].

PharmGKB - The Pharmacogenomics Knowledge Base [97] The Pharmacogenomics Knowledge Base (PharmGKB) is specialized on the knowledge about pharmacogenes, genes that are involved in modulating drug response. Genes are classified as

[4] http://www.nlm.nih.gov/research/umls/

[5] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene

pharmacogenes because they are (i) involved in the pharmacokinetics of a drug (how the drug is absorbed, distributed, metabolized and eliminated) or (ii) the pharmacodynamics of a drug, i. e. how the drug acts on its target and its mechanisms of action. Hence, it covers human gene-disease associations less broadly but was found to be complementary to the other sources as it contains some gene-disease associations not present in the other repositories. The data[6] was downloaded on June, 6th 2009. We included 1.772 associations for 79 distinct genes and 261 distinct diseases. Dependent on the original classification in PharmGKB, the associations were either classified as *Marker* or as *RegulatoryModification* in the gene-disease association ontology [83].

CTD - The Comparative Toxicogenomics Database [129] The Comparative Toxicogenomics Database (CTD) contains manually curated information about gene-disease relationships with focus on understanding the effects of environmental chemicals on human health. The data[7] was downloaded on June, 2nd 2009. CTD is a large gene-disease association repository, but also consists partly of relations originating from OMIM enriched with additional information. All available gene-disease associations and all cross links to PubMed were kept. In total, CTD data provided 6.469 association for 2.702 distinct diseases and 3.345 distinct genes. Dependent on the original classification in CTD, the associations were either classified as *Marker* or as *Therapeutic* in the gene-disease association ontology [83].

ALL - a unified database [83] As already motivated, every database is restricted due to its specific focus. However, to understand complex diseases a unified view is needed. This increases important properties such as coverage as well as reliability of the database. ALL integrates all data sources mentioned before plus the LHGDN (see next Section 4.3) and thus represents a huge gene-disease association database. It is composed of four data sources that are curated by humans and one text mining derived resource. If only the curated sources are merged together without using the LHGDN, then we call this unified database CURATED. The different data sources were integrated through the gene-disease association ontology (see Figure A.1). The main goal of the ontology was to harmonize the different views for gene-disease relationships used in the different data sources.

Regarding the different controlled vocabularies of the data sources, CURATED and ALL use the same controlled vocabulary for gene entities. All sources, except UniProt, use Entrez Gene [182] identifiers. The mapping from UniProt to Entrez Gene is complete enough to use one controlled vocabulary for genes. However, this does not apply to the disease vocabulary. Some sources use the disease vocabulary OMIM Morbid Map provided provided by OMIM (OMIM, UniProt), while others rely on Medical Subject Headings (MeSH) (PharmGKB, LHGDN). An exception is the CTD database that uses both disease vocabularies. In general, the classification of phenotypes is highly complex and OMIM uses a more fine grained vocabulary than MeSH. The mapping is non-trivial and subject of

[6] http://www.pharmgkb.org/resources/downloads_and_web_services.jsp
[7] http://ctd.mdibl.org/downloads/

current research (see e. g. [117]). It is currently solved with an iterative procedure. First, all OMIM disease terms are mapped to MeSH via the UMLS Metathesaurus[8]. Then, the remaining terms are mapped with a combination of string similarity methods and careful manual inspection. Details can be found in [83]. By using this strategy, approximately 50% of the OMIM Morbid Map vocabulary can be mapped to MeSH. Therefore, CURATED and ALL are composed of two controlled disease vocabularies.

4.3 Literature-derived Human Gene-Disease Network (LHGDN)

In Chapter 2, two CRF models for extracting facts from text were introduced. In this section *Text2SemRel* is trained on the complete in-house generated corpus from Section 2.4.3. In particular, the cascaded CRF was chosen, since it yielded a better performance than the one-step CRF (see Chapter 2, Section 2.4.3). Afterwards, *Text2SemRel* was applied to the whole Entrez Gene's GeneRIF (Gene Reference Into Function) database. The GeneRIF database represents a rapidly growing knowledge repository and consists of high-quality phrases created or reviewed by MeSH indexers. Hereby, the phrases refer to a particular gene in the Entrez Gene database and describe its function in a concise phrase. Using this textual repository for text mining has recently gained increasing attention, due to the high quality of the provided textual data in the GeneRIF database. The LHGDN was created based on a GeneRIF version from March 31st, 2009, consisting of 414.241 phrases. These phrases were further restricted to the organism Homo Sapiens, which resulted in a total of 178.004 phrases.

Text2SemRel needed about six hours on a standard Linux PC with an Intel Pentium IV processor 3.2 GHz to extract all facts found in the GeneRIF database. The extracted facts were normalized to URIs and stored in form of a RDF graph[9]. To be compliant with the linked data principles, the entity mentions were normalized to already existing URIs using the Bio2RDF[10] namespaces. Bio2RDF aims to transform database silos in the biomedical domain into a platform for distributed biological knowledge discovery [17].

The GeneRIF database is an encyclopaedic-style text collection (see Chapter 2) where the textual phrases refer to a particular gene. This makes the normalization of gene names to URIs trivial. The identified disease mentions in the text were normalized with the sliding window heuristic presented in Section 2.2.5. Bio2RDF links to the MeSH thesaurus, which was used as controlled vocabulary for normalization. Besides the MeSH heading, which represents the common official name for a concept, alternative entries were used as synonyms. MeSH is divided into several branches and we restricted to the C branch of MeSH, which stores disease concepts. Note that this normalization procedure is not perfect and thus false positive facts are also extracted. E. g. the CRF model often

[8] http://www.nlm.nih.gov/research/umls/

[9] http://www.dbs.ifi.lmu.de/~bundschu/LHGDN.html

[10] http://bio2rdf.org/

tagged 'oxidative stress' as disease mention, which is then mapped with the sliding window heuristic to the ontology entry stress. Even though the MeSH vocabulary provides a good coverage of synonyms, it is far from being perfect. This represents another source of errors in the normalization strategy. As a concrete example, consider the disease mention 'mammary tumors', a synonym for the MeSH concept 'Breast Neoplasms'. The phrase is not part of the official synonym list of the MeSH entry 'Breast Neoplasms', while e. g. 'mammary neoplasms' is. As a consequence, the heuristic can only map 'mammary tumors' to 'Neoplasms', because the term 'tumors' is in the synonym list for 'Neoplasms'. Indeed, this introduces some undesired impreciseness, but the mapping is actually not wrong as the concept 'Breast Neoplasms' stands in a is-a relationship with 'Neoplasms'.

In contrast to the gene-disease associations provided by the just introduced curated state of the art databases, gene-disease associations in the LHGDN are classified into several biomolecular conditions. The biomolecular conditions are describing a wide variety of molecular conditions, ranging from genetic to transcriptional and phosphorylation events. In particular, we defined the following conditions that can hold between genes and diseases: *altered expression, genetic variation, regulatory modification, any relation and negative associations* (for more details, see Section 2.4.3 and the provided annotation guidelines available online[11]). We believe that the chosen biomolecular conditions represent important, valuable conditions for a better understanding of the disease itself and its underlying mechanisms. E. g. new reported mutations expressed by *genetic variation* facts, might give new insights about the etiology of a disease. Facts about *regulatory modifications* of a gene in a disease might give a better understanding about regulatory mechanisms in diseases. *Altered expression* levels, e. g. increased protein/gene levels in a disease point to potential new biomarkers. But also less specific facts such as *any* relations between genes and diseases or *negative associations* represent valuable sources of information and might assist researchers in designing new experiments. However, as shown in Section 2.3, the full power of knowledge discovery unfolds with an interactive, graphical user interface in combination with simple filtering and keyword search functionalities.

In total, the LHGDN provides currently 59.342 positive statements about gene-disease associations for 1.850 diseases and 6.154 distinct genes. This represents a much larger information space than currently provided by other curated state of the art databases. The LHGDN is publicly available[12]. Facts in the LHGDN are encoded in RDF using N-triples[13]. Reification is used to add further statements about a found gene-disease association such as the PubMed article, where the association was found or the type of relation that is predicted by *Text2SemRel*. By using existing URI's from the Bio2RDF project we are able to link genes and diseases to additional information, such as the chromosomal location of genes or additional annotations such as Gene Ontology annotations.

A previous version of the LHGDN extracted from a GeneRIF version from August 2007 is available online as supplementary data[14]. This old version consists of 34.758 statements

[11] http://www.biomedcentral.com/1471-2105/9/207/additional/

[12] www.dbs.ifi.lmu.de/~bundschu/LHGDN.html

[13] http://www.w3.org/TR/rdf-testcases/#ntriples

[14] http://www.biomedcentral.com/1471-2105/9/207/additional/

about gene-disease associations for 4.939 unique genes and 1.745 unique diseases [34].

The More is not Always the Better: the Role of the GeneRIF Database. Criticism could be expressed against analyzing GeneRIF sentences rather than making use of the enormous information available from original publications. However, GeneRIF phrases are of high quality, as each phrase is either created or reviewed by professional MeSH indexers. Moreover, the number of available sentences is growing rapidly[15]. Thus, analyzing GeneRIFs might be advantageous compared to a full text analysis, as noise and unnecessary text is already filtered out. Some existing related work underscore this hypothesis. E. g. , [164] set up an annotation tool for microarray results based on two literature databases: PubMed and GeneRIF. They conclude that a number of benefits resulted from using GeneRIFs, including a significant decrease of false positives as well as an apparent reduction of search time. A further study highlighting advantages resulting from mining GeneRIFs is the work of [124]. Building upon our here presented work [34], [146] recently compared gene-disease associations mined from the GeneRIF database with OMIM, concluding that the mined gene-disease associations are far more complete than the knowledge available in OMIM.

To summarize, the GeneRIF database is an excellent source for extracting information in the biomedical domain. Mining this database comes with the following key advantages:

- GeneRIF phrases are of high quality as they are either created or reviewed by professional MeSH indexers.
- The number of phrases to mine is growing rapidly.
- GeneRIF phrases are making statements about particular genes, thus the key entity is already given. The particular difficult task of gene mention normalization can be omitted or at least reduced (see e. g. the BioCreAtIvE 2 evaluation on gene mention normalization [139]).
- Despite it's small size, when compared to the whole PubMed database, our experimental analysis in this chapter reveals that a lot of knowledge currently buried in the literature, can be unlocked when mining the GeneRIF text collection (see Figure 4.1).

4.4 Analysis of the LHGDN

In the last section the different databases were merged to a single, huge gene-disease association repository. This section is devoted to analyze the LHGDN by means of size, quality and large-scale properties of the discovered facts. We start with a quantitative analysis, and compare the LHGDN to other existing state of the art gene-disease associations databases. Afterwards, we would like to get a notion about the quality of the extracted network by

[15] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/stats/

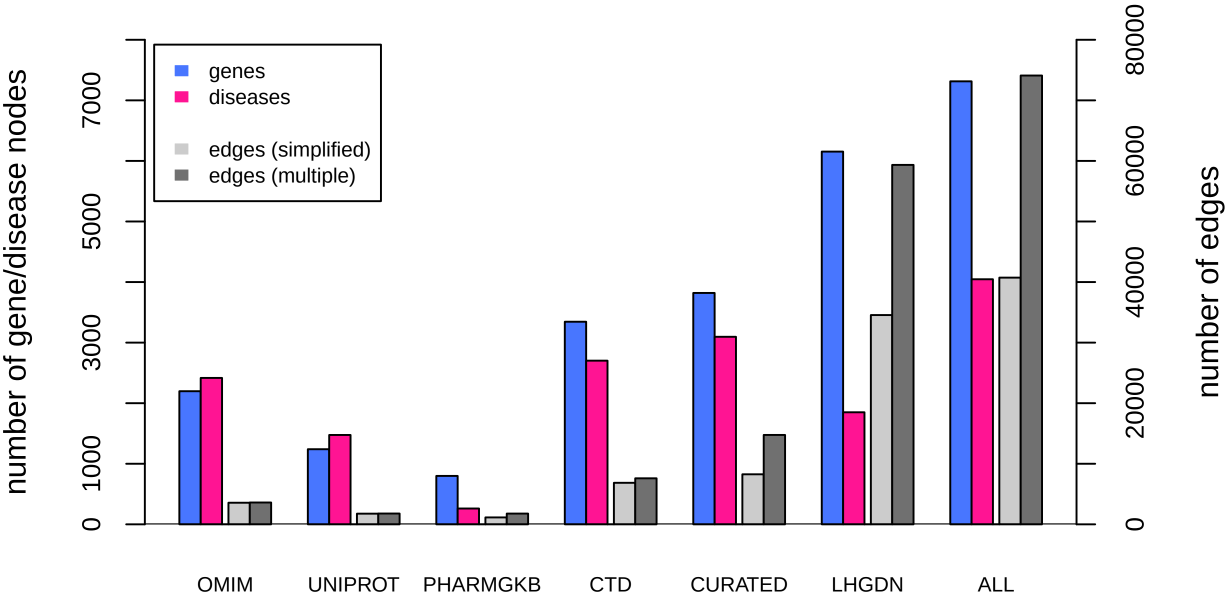
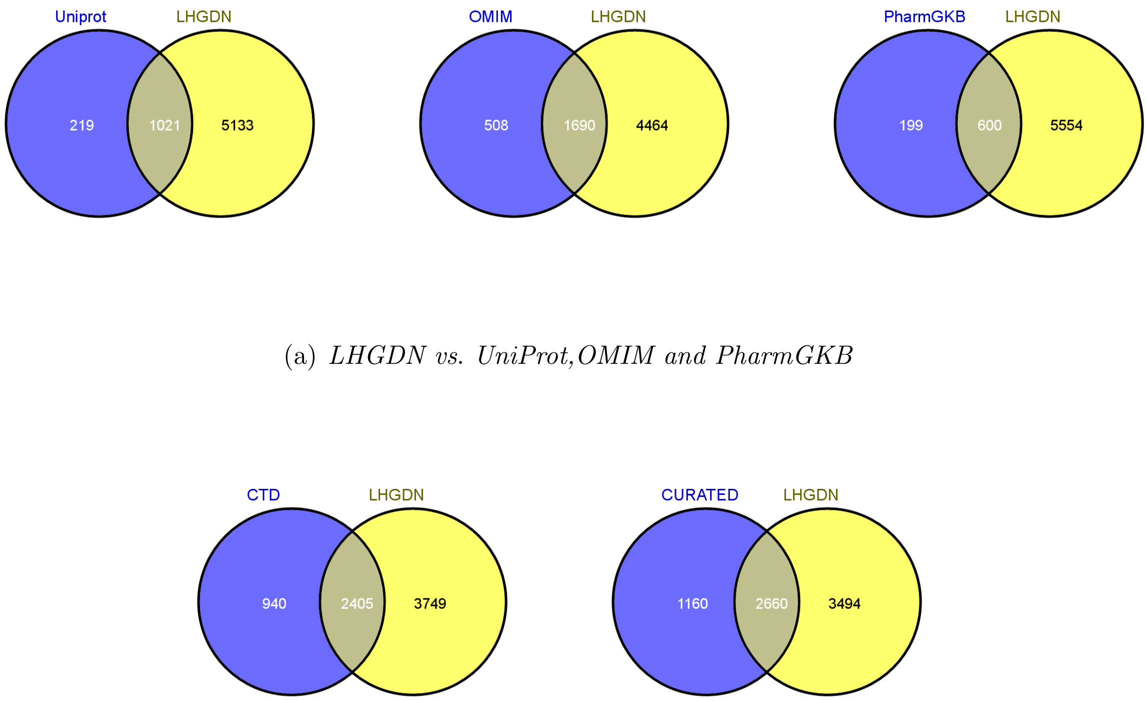


Figure 4.1: **Number of genes, diseases and edges in the databases under consideration.** Edges (simplified) represent unique gene-disease associations, while multiple denotes that an association is counted every time it is mentioned in a database.

comparing disease gene properties of the various databases. The section is completed with an analysis of the large-scale properties of the literature derived network.

4.4.1 Quantitative Analysis

Figure 4.1 plots the distinct number of genes, diseases and the number of associations between genes and diseases with respect to the single data repositories. PharmGKB represents the smallest repository, because it discusses gene-disease associations only in the context of pharmacogenomics. Regarding the number of gene-disease associations (referred to as edges in Figure 4.1), PharmGKB is followed by UniProt which stores only genetic variations. OMIM covers mainly Mendelian diseases, where the phenotype is inherited or controlled by a single gene. Only recently, OMIM started to include diseases with complex traits and genetic mutations that confer susceptibility to a disease [134]. The largest curated database, CTD, is primary specializing in storing gene-disease associations in the context of environmental factors. CTD has some associations stored several times, since it uses partly the facts stored in OMIM [129]. CURATED, the unification of all curated databases already represents a large association repository. However, the LHGDN covers a broad spectrum of different gene-disease associations and contains by far the most gene-disease associations. When compared to the other databases, the relative small size of disease mentions, is notable. This is likely to originate from using MeSH as a controlled disease vocabulary, which is not as fine-grained as the OMIM Morbid Map. Another reason might also be that GeneRIF authors do not use complex phenotype concepts when stating



(a) *LHGDN vs. UniProt,OMIM and PharmGKB*

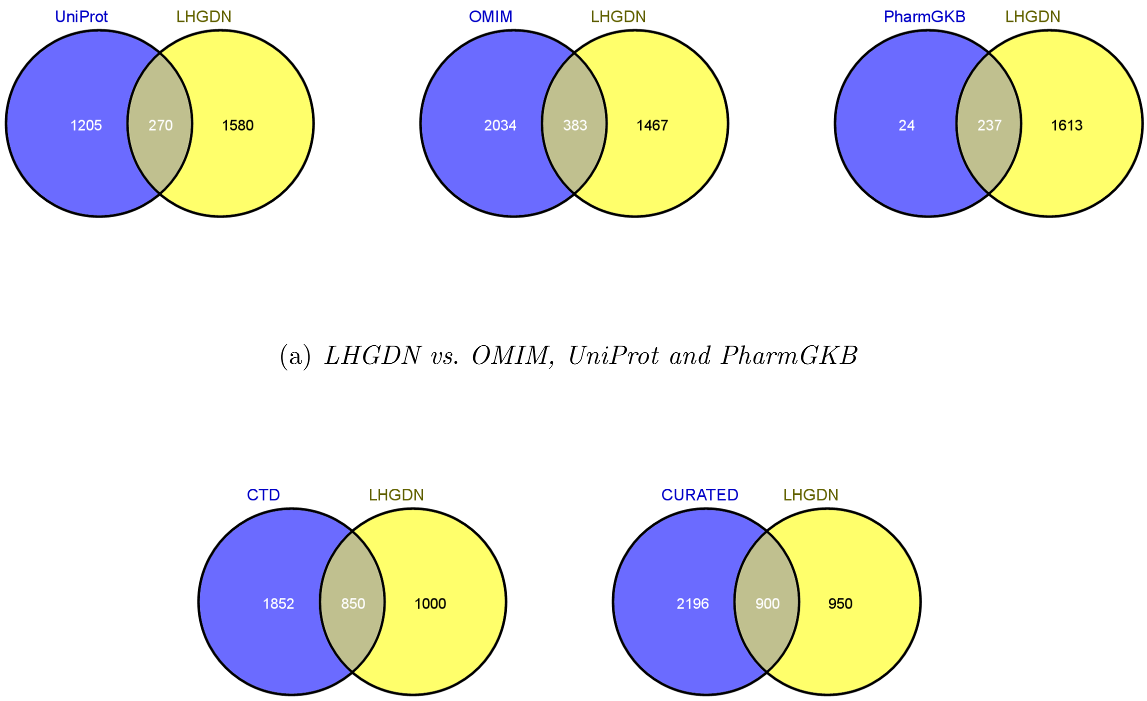
(b) *LHGDN vs. CTD and Curated*

Figure 4.2: **Comparison of gene coverage of LHGDN with curated gene-disease databases.**

the main results of a study. Fairly speaking, despite the huge information space that the LHGDN provides, it contains, of course, also false positive associations that are introduced by the text mining approach.

Figure 4.2 gives an idea about the amount of overlapping genes of the various databases with the LHGDN. The last Venn diagram Curated vs. LHGDN shows that despite a large overlap of genes that can be found in both, the curated databases and the literature, the unique sets of genes are, in both cases, still large. From a total of 7314 genes, approx. 36% can be found in both, the literature derived database and the curated databases. The fraction of unique genes originating from curated databases is still about 16%, while the fraction of genes that can only be found in the LHGDN represents the largest fraction with about 48%.

As already mentioned in Section 4.2, except the UniProt repository, all data repositories use gene identifiers from Entrez Gene. There exist well defined mappings between Entrez Gene and UniProt, which makes it straightforward to compare the gene coverage between



(a) *LHGDN vs. OMIM, UniProt and PharmGKB*

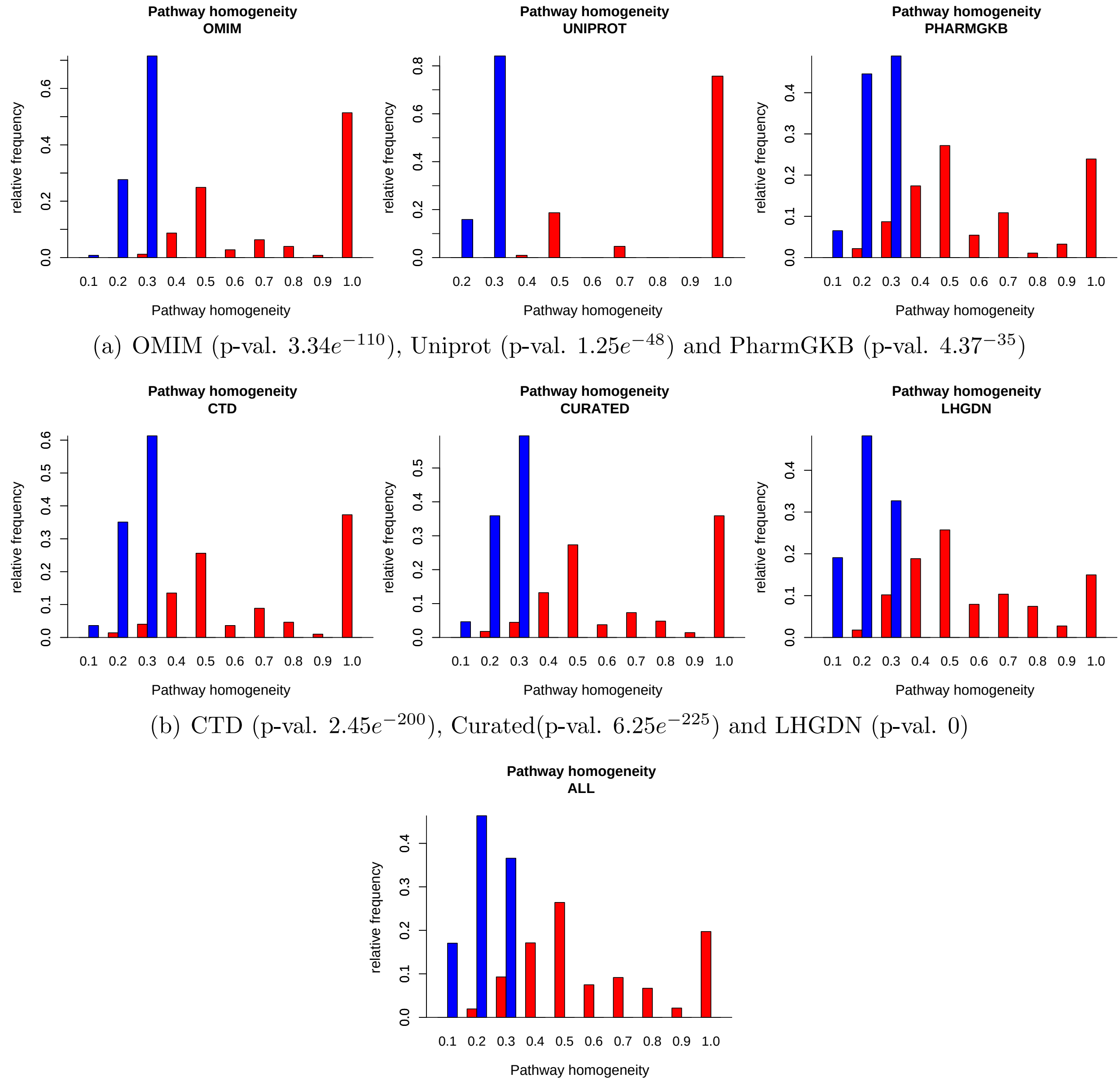
(b) *LHGDN vs. PharmGKB, CTD and Curated*

Figure 4.3: **Comparison of disease coverage of LHGDN against curated gene-disease databases.** Note that the disease mapping from the different databases are far from being perfect.

the databases. For diseases, however, this is not the case and the mapping is far from being perfect. Thus, it is difficult to judge the true differences of disease coverage between the LHGDN and the other databases. However, it can be seen that even though CURATED has far more disease terms than the LHGDN (Figure 4.3), the LHGDN has still many disease terms that are not part of the CURATED database (approx. 950).

4.4.2 Disease Gene Property Analysis

The characterization of disease genes is a major focus of current research in translational bioinformatics. Deeper insights into the properties of disease genes has yielded a better understanding of diseases and has the great potential to personalize medicine. Following the notion of [112], we refer to disease genes as those genes that are either involved in the causation of, or associated with a disease. The analysis of the characteristics of disease



(a) OMIM (p-val. $3.34e^{-110}$), Uniprot (p-val. $1.25e^{-48}$) and PharmGKB (p-val. 4.37^{-35})

(b) CTD (p-val. $2.45e^{-200}$), Curated(p-val. $6.25e^{-225}$) and LHGDN (p-val. 0)

(c) ALL (p-val. 0)

Figure 4.4: **Pathway homogeneity of the different databases.** The p-values indicate the probability that the two distributions (random and observed) are the same. The x-axis represents the homogeneity values and the y-axis shows the fraction of homogeneity values obtained.

genes vs. non-disease genes has lead to the development of disease classifiers, which, in turn, can be used to prioritize genes that are potentially involved in a disease. The recent review of [112] gives a general overview over disease gene properties studied currently in the literature.

This section is devoted to use some of the existing known disease gene properties to compute statistics that characterize the investigated databases. To the best of our knowledge, up to now, only OMIM was used to characterize human disease genes [88], while other repositories for human gene-disease associations have not been subject of consideration. By comparing the investigated databases with statistics about disease properties, we would like to get insights if there are any significant differences between those databases. Intuitively, we would expect that statistics about disease gene properties vary from database to database, because every database has its own focus and thus stores different kinds of gene-disease relations. In particular, two properties of disease genes are used that have been shown to represent intrinsic properties in large-scale gene-disease networks: (i) homogeneity of the function of genes related to a disease and (ii) homogeneity of pathway annotations of genes that are related to a disease. Note that while the first homogeneity measure has already been studied in the literature (but only for OMIM) [88], the second measure based on pathways has not been reported in the literature so far. The notion behind these two properties is that genes are most likely to share common annotation of gene functions as well as pathways involved in the same disease. This hypothesis has been validated in previous work for gene function homogeneity by using the OMIM database [88].

To measure the gene function homogeneity, the annotations of the genes originate from the Gene Ontology (GO) [93]. The Gene Ontology provides a controlled vocabulary of terms to describe gene functions and other characteristics of gene products. GO is composed of three branches for gene annotations (molecular function, cellular component, and biological process). Each of these branches is hierarchically organized by is-a relationships. Additional information for gene product annotations is provided as well. E. g. for each annotation, a link to the original PubMed publication that gives evidence for the annotation is provided, together with a classification of the experimental evidence (for more information see [93]). Each gene can have multiple GO annotations. We used a GO version downloaded on October 1st, 2009. For the second measurement, the pathway homogeneity, the Reactome [128] database was used to retrieve pathway annotations. The downloaded version is from November 1st, 2009.

Following the notion of [88], the homogeneity of a disease d_i (H_{d_i}) is defined as the maximum fraction of genes in a disease d_i that share the same annotation from a controlled vocabulary $\mathbf{V}$:

$$H_{d_i}(\mathbf{V}) = \max_{j \in \mathbf{V}} \frac{n_i^j}{n_i}, \tag{4.1}$$

where n_i is the number of genes in a disease d_i that have any annotations with respect to the predefined vocabulary $\mathbf{V}$, and n_i^j is the number of genes that have the specific annotation j. Note that H_{d_i} is only computed when $n_i \geq 2$.

The homogeneity analysis now works as follows: We compute homogeneity values separately for each of the three branches of GO (molecular function, biological process, and cellular component) as well as for the pathway annotations obtained from Reactome. Afterwards, the obtained values are compared to random control trials for each of the databases. The random control trials are computed by randomly choosing the same number of dis-

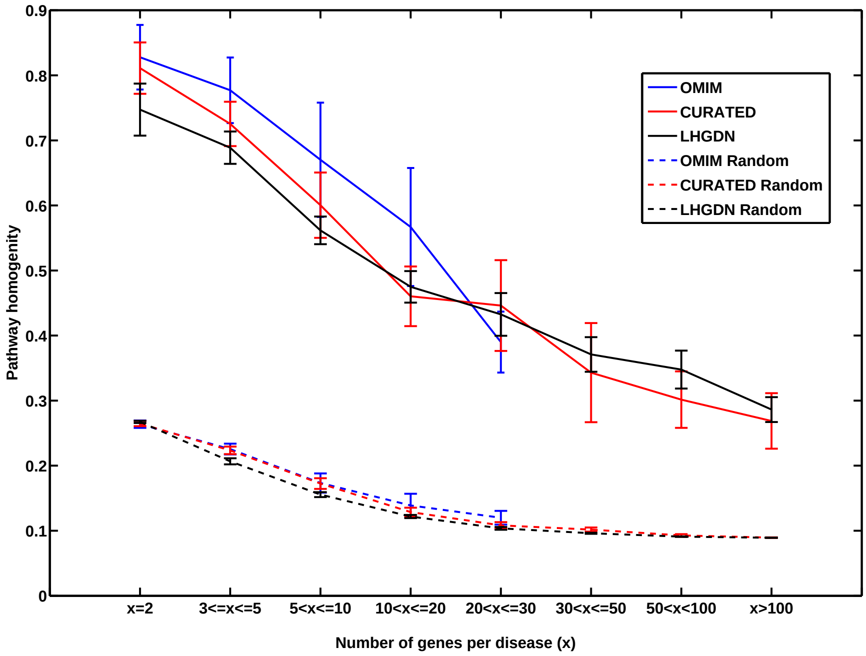


Figure 4.5: **Pathway homogeneity plot that shows the development of homogeneity with respect to the number of genes per disease.** Error bars represent the standard error of the mean (95% confidence interval.)

ease genes for a disease and taking the resulting random annotations for computing the homogeneity. The random trials are repeated a sufficient number of times in order to reach statistical significance (10^5 times). Afterwards, a two-sample Kolmogorov-Smirnov (KS) test at the 95% significance level is performed for each database. The null hypothesis is that the random control trial and the homogeneity values obtained for a single database are drawn from the same distribution. P-values are computed that give us the probability for seeing such a test statistic at least as extreme as the one actually observed, assuming that the null hypothesis is true. Figure 4.4 shows the results for the pathway homogeneity analysis. The blue bars indicate the distribution obtained from the random control trial and the red bars are obtained by taking the genes and their corresponding annotations from the different databases. The distributions of homogeneity values for all databases are significantly different from the random control trials (see p-values in the figure). With increasing size of samples (i. e. with increasing size of the databases), the probability that the random control trials and the observed distribution are the same can even be ruled out. The figure shows that the two data repositories, OMIM and UniProt, that store solely information about genetic variations or mutations in diseases have the largest fraction of perfect homogeneity values (OMIM$\sim$ 0.52, UniProt$\sim$ 0.77). This fraction of perfect homogeneity matches drops for the other databases (PharmGKB$\sim$ 0.25, CTD$\sim$ 0.38, CURATED$\sim$ 0.39, LHGDN$\sim$ 0.16, ALL$\sim$ 0.20). However, it can be easily seen that the homogeneity computation chosen in [88] is sensitive to the number of genes that a disease has. See also the random curves in Figure 4.5. Thus, the homogeneity values of the various databases cannot be directly compared, since the number of genes per disease varies tremendously between the single databases.

To further understand this behavior and in order to be able to compare homogeneity across various databases, we group diseases that share the same number of genes within a given interval and compute the average homogeneity for the interval (see Figure 4.5). The intervals were chosen empirically to ensure that the small databases (OMIM, UniProt and PharmGKB) have a sufficient number of samples inside each interval such that the standard error of the mean is minimized. For the sake of clarity, we only show OMIM, Curated, and the LHGDN. A general trend that can be seen is that if the number of genes that are involved in a disease increases, the homogeneity decreases dramatically for all data repositories. This means that the more complex a disease is, the more pathways are involved in a disease. The same trend still holds, when homogeneity values are computed with GO annotations (see Figure A.5). For intervals with a small number of genes per disease (up to the interval $10 < x \leq 20$), OMIM has the largest homogeneity values, followed by Curated and the LHGDN. Therefore, the large difference of relative frequencies of perfect homogeneity matches in Figure 4.4 between the databases that store solely genetic variations and the remaining data sources can be mainly explained by the small number of genes per disease in these databases (OMIM, UniProt). It is likely that missing data also play a role here. Another small effect might also be due to the different focuses of the databases, which in turn reflect different proportions of homogeneity values. E. g. the major focus of PharmGKB is to collect knowledge about the impact of human genetic variations on drug response in diseases. Thus, the signature of homogeneity values in a disease might vary when compared to OMIM, which stores solely information about inherited genetic disorders. The difference in homogeneity values for intervals with a small number of genes between CURATED and the LHGDN can likely be explained by the fact that the text mining derived database introduces some noise. However, as the number of genes per disease grows, there is no distinguishable difference between the databases anymore.

In Appendix A, we present homogeneity plots as supporting information by using the Gene Ontology for gene annotations. The results are quite similar to the pathway homogeneity. Figure A.2 shows results for biological process, Figure A.3 for the GO branch cellular component and Figure A.4 uses the GO branch molecular function for gene annotation. Finally, Figure A.5 shows the development of the GO homogeneity values with respect to the number of genes per disease.

To summarize the findings in this section: The homogeneity analysis shows that diseases tend to share similar functional annotations (in terms of pathway annotations and GO annotations). This hypothesis holds across different databases. It can also be seen that the trend to share similar functional annotations differs heavily from random control trials. As diseases become more complex and the number of genes per disease increases, the homogeneity values drop significantly (see Figure 4.5 and Figure A.5).

4.4.3 Large-scale Properties of Discovered Facts

In the last two sections the LHGDN was compared with other existing curated databases. Here we analyze properties that emerge from the nature of the network: the literature. The LHGDN is investigated regarding how facts are published. We are looking for intrinsic patterns how facts and their components are distributed to get hints about the dynamics and the underlying processes of the evolution of this literature-derived network. This will provide insights about how the knowledge of gene-disease associations expands and will help to predict the future evolution of the network.

We hereby make use of findings and methods from the emerging field of network science [3]. Network science studies network representations from heterogeneous domains such as e. g. biological networks, social networks and information networks. Traditionally these networks have been modeled as random graphs [31]. Informally speaking, networks in random graph theory are modeled by randomly placing links between nodes. As a consequence, it is extremely unlikely to find nodes whose number of links is significantly different from the mean. Recently, researchers became aware of the fact that for many real-world networks such as the web, the true link distributions of nodes cannot be explained or modeled with random graphs. As a major result, researchers developed the theory of scale-free networks [3], which is able to explain and model critical phenomena in these complex networks such as the occurrence of hubs. Hereby, hubs are nodes that are dominating the network by having a huge number of links that greatly exceeds the average. In these networks the distribution of links follows a power law and it has been shown that a number of real-world networks exhibit this behavior: e. g. the web, protein-protein interaction networks, citation networks and social networks. Traditional network analyses typically study characteristics such as the average in- and out-degree distributions of nodes (i. e. the number of incoming and outgoing links a node has), average path lengths between nodes or clustering coefficients. See [3] for more details about further examples and properties in these networks.

Distribution of Entity and Fact Mentions in the LHGDN

In the following analysis, we concentrate on distributions of specific features in a literature-derived network. From a graph or network perspective, the LHGDN can be regarded as a weighted graph. A weighted graph is a graph, where each edge e is associated with a weight w_e. Edges in the LHGDN are created between disease and gene nodes, constituting a fact. The weight w_e indicates how often a gene-disease association is mentioned in the literature. Thus, each time a specific gene-disease association is discussed in a publication, w_e increases. In particular, the following quantities are investigated:

- The distribution of the number of times a specific entity (a specific node n) is mentioned in the literature (here genes and diseases). This quantity can be expressed by $\sum_{e \in E_n} w_e$, where E_n is the set of edges of node n.

- The distribution of the number of times a fact is mentioned in the literature (here a gene-disease association). Thus, this quantity is simply the distribution of the edge weights in the LHGDN.

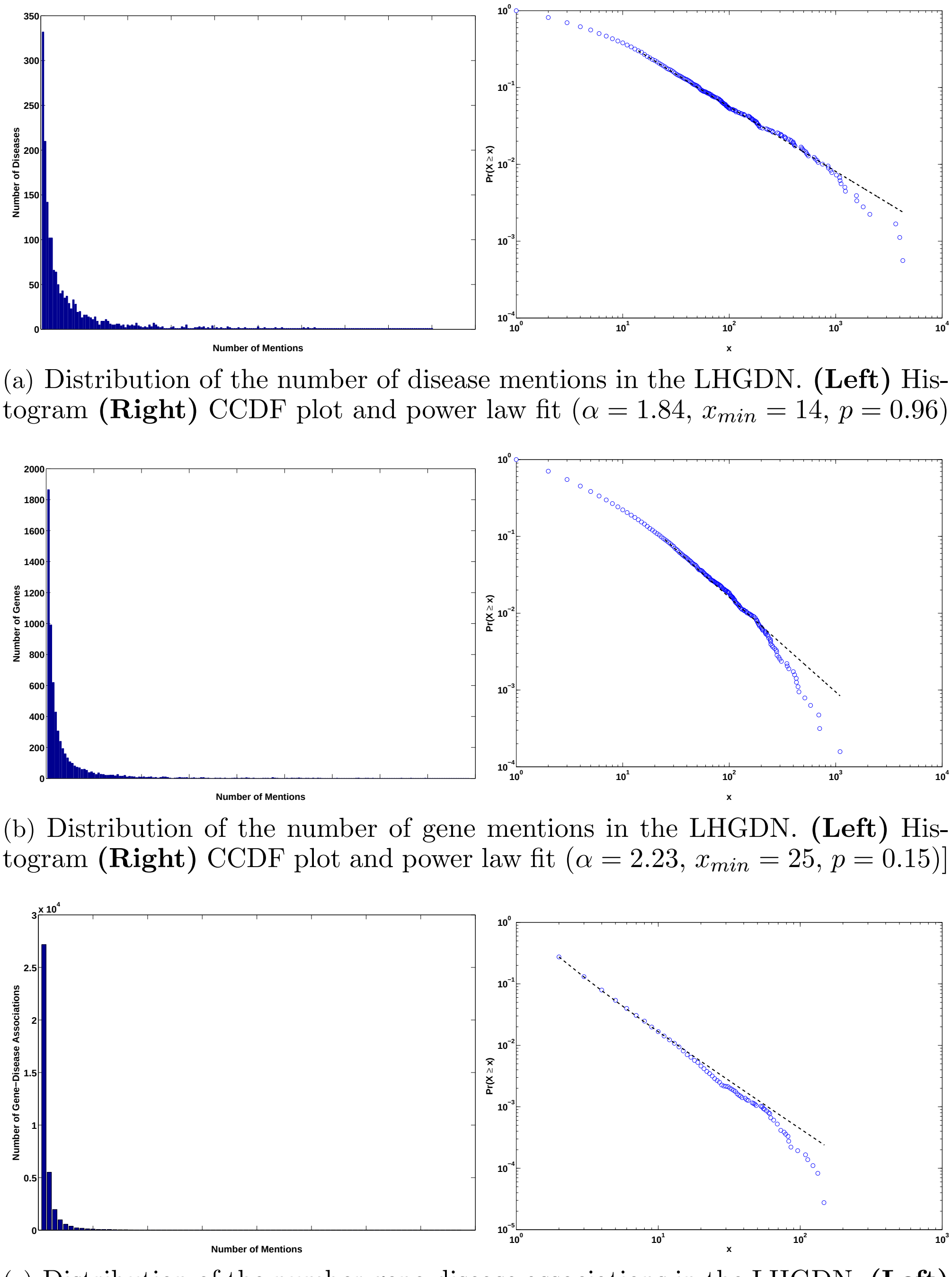
The analysis of these distributions will give us an idea about how gene-disease associations are discussed by the researchers. It will also give insights whether associations, genes or diseases are discussed more uniformly in the literature or whether some associations, genes or diseases are more heavily discussed than others.

To get a first idea about the distributions of interest (i. e. gene-disease associations mentions, gene mentions, and disease mentions), we investigate the mean of the number of times they are mentioned. In average, every fact about a gene-disease association is discussed $\bar{f} \sim 1.87$ times, each gene is mentioned $\bar{g} \sim 11.17$ times and each disease is discussed approx. $\bar{d} \sim 39.8$ times. Inspecting the distributions more carefully, it becomes clear that the observed empirical quantities do not cluster around a typical value. For all three distributions there are values that have huge deviations from the mean. E. g. , regarding gene-disease associations, one can find approx. 200 values that are about a factor of 10 away from the mean, with a maximum value deviating with a factor of about 84 from the mean. The same holds for the entities of interest, where we can find nearly 100 gene values (respectively 33 disease values) that deviate with a factor of 10 from the mean. The maximum gene value deviates with a factor of nearly 100, respectively 110 for disease entities. This behavior makes a normal distribution, where a factor of two from the mean is already a very rare event, extremely unlikely. The histograms on the left side of Figure 4.6, with their long tails, exemplify this behavior. Note that a long or a heavy tail is a property of power law distributions. Indeed, from a first visual inspection, the histograms look like they could follow a kind of power law distribution. According to [59], a quantity x is said to have a power law distribution when

$$p(x) \propto x^{-\alpha}, \tag{4.2}$$

where α is called the scaling exponent. This exponent typically lies within the range of $2 < x < 3$, as observed in many empirical quantities [3]. This distribution reveals that the quantities are scale-free in the sense that some values seem to have an unreasonable high quantity. In the majority of cases, a power law does not hold for all values of x. Instead, it often holds only for a lower bound x_{min}. In what follows, we are investigating if statistical support for this hypothesis can be found.

A result from the study of heavy-tailed distributions is that a quantity of interest X will asymptotically show the behavior of a straight line in a log-log plot of the Complementary Cumulative Distribution Function (CCDF) $P(X > x)$ [137]. This represents a simple empirical test for whether X is following a power law or not [137]. Note that the CCDF can be simply computed by $1 - P(X \leq x)$, with $P(X \leq x)$ being the Cumulative Distribution Function (CDF). The blue dots in the plots on the right-hand side of Figure 4.6 show the CCDF of the empirical quantities in a log-log plot. It can be seen that the empirical quantities indeed seem to follow a straight line. However, validating power laws in empirical data is particularly difficult and is an active field of research [59]. Recently, [59] analyzed 24 data sets that were assumed to follow a power law behavior in previous work. For



(a) Distribution of the number of disease mentions in the LHGDN. **(Left)** Histogram **(Right)** CCDF plot and power law fit ($\alpha = 1.84$, $x_{min} = 14$, $p = 0.96$)

(b) Distribution of the number of gene mentions in the LHGDN. **(Left)** Histogram **(Right)** CCDF plot and power law fit ($\alpha = 2.23$, $x_{min} = 25$, $p = 0.15$)]

(c) Distribution of the number gene-disease associations in the LHGDN. **(Left)** Histogram **(Right)** CCDF plot and power law fit ($\alpha = 2.55$, $x_{min} = 2$, $p = 0.10$)

Figure 4.6: **Distributions of the number of times entities and facts are mentioned in LHGDN.** The left figures show the histogram of the empirical quantity, while the right figures show a log-log plot of the complementary cumulative distribution function.

some of the data sets no evidence for such a distribution could be found and the author's showed that sometimes wrong statistical assumptions were made (see [59] for more details). Therefore, we decided to follow the rigorous evaluation procedure proposed in [59] and use the software that the authors released with the paper[16]. In what follows, this evaluation procedure is used to test if a power law in our empirical quantities of interest is a reasonable assumption.

The first task is to fit a power law to the observed data. This is often done in the literature by taking the logarithm of the power law function $\log p(x) = -\alpha \log x$ and fitting a least-square linear regression to the data. The slope of the function is then interpreted as the scaling exponent α. However, as [59] shows, this estimate is often biased and yields inconsistent answers. Instead, Maximum Likelihood Estimation (MLE) can be used to estimate α more properly. Determining the lower bound x_{min} is done with statistical testing. The interested reader is referred to [59] for mathematical details.

The estimated values α and x_{min} are given in Figure 4.6, (a),(b),(c), respectively. The black lines in the right figures represent the power law fit to the data (see Figure 4.6, (a),(b),(c) right-hand side). It can be seen that for large values of x, all three empirical quantities, while still being a good fit, start to deviate from the power law estimation.

Up to now, we determined the best values that fit the data. In order to be able to judge whether a power law is a plausible hypothesis or not, a goodness-of-fit test has to be accomplished . The goodness-of-fit test proposed in [59] works as follows: First, a power law is fitted to the observed data in order to determine α and x_{min}. Afterwards, a large number of synthetic data sets, which are power-law distributed, are generated with the parameters determined from the fit to the empirical data. Each artificially generated data set is then fit to its own power law. In this way, statistical fluctuations that arise from the sampling process can be accessed. For each generated data set and the corresponding fitted power law, the distance between the two distributions is computed, which can be done with the well-known Kolmogorov-Smirnov distance. To access statistical significance, the p-value is defined to be the fraction of the number of times the obtained synthetic distances are larger than the empirical distance. Therefore, if the distance of the synthetic data sets is almost always larger than the empirical distance, the p-value is close to one. In this case, the difference between the fitted power law and the empirical data can be classified as statistical fluctuations. The number of synthetic samples that have to be drawn and the cut-off value for the p-value are determined by statistical analyses in [59]. It is concluded that in order to obtain p-values that deviate $\epsilon = 0.01$ from the true value, a good choice for the number of samples is $n = 2500$. The cut-off value is chosen very conservatively and if the p-value exceeds $p = 0.1$ than the hypothesis that the observed empirical quantity follows a power law is plausible.

Following the proposed goodness-of-fit test, we observe the following p-values: $p = 0.96$ for the number of disease mentions ($x_{min} = 14$), $p = 0.15$ for the number of gene mentions ($x_{min} = 25$), and $p = 0.10$ for the occurrence of gene-disease associations ($x_{min} = 2$). Thus, all three empirical quantities observed in the current status of the LHGDN are plausible

[16] http://tuvalu.santafe.edu/~aaronc/powerlaws/

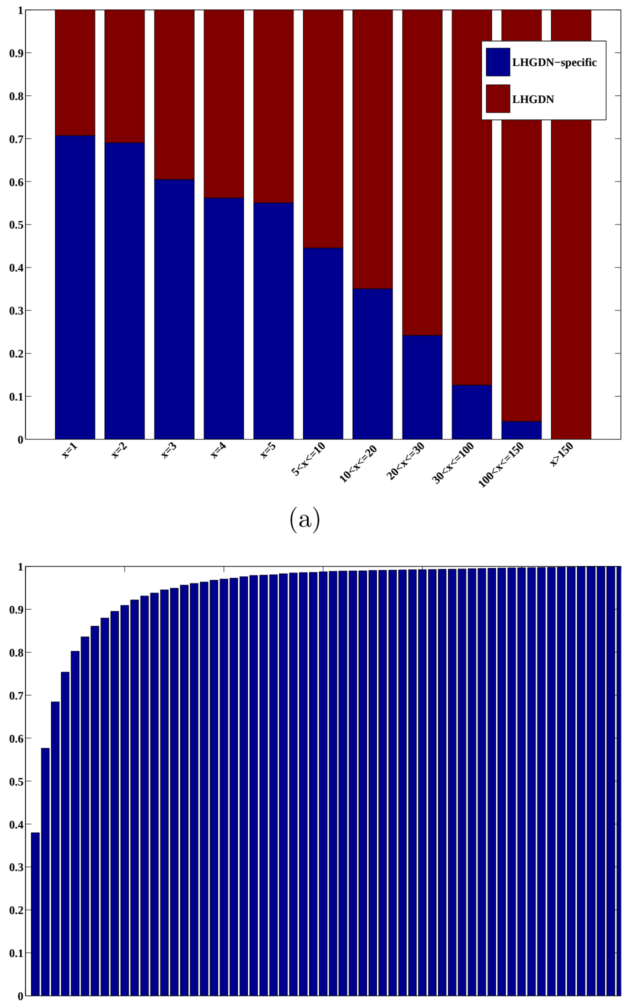
to be drawn from a power law distributions.

There are different theories about the underlying generative processes that can create quantities with a power law behavior. Two classical mechanisms that can produce power law distributions together are growth and preferential attachment. In general, preferential attachment is a process that models cumulative advantage, i. e. a variable that already has a high value is more likely to grow in the future than a variable that is small. This behavior is also known as 'the rich get richer' effect [3]. Growth and preferential attachment might be reasonable mechanisms for generating the power law behavior of the observed quantities in the LHGDN as well. Once a gene-disease association is added to the net, other researchers will recognize this and will study the gene-disease associations with regard to their research focus. Other gene-disease associations will be discussed more controversial and thus be also subject of growth in the LHGDN. Some associations might be highly complex and thus attracting the interest of many researchers. The same holds for genes and diseases. Once a new disease gene is discovered, it will be investigated if it might also be a disease gene in other similar diseases. Interestingly, gene and disease quantities have a rather large lower bound x_{min} (Diseases: $x_{min} = 14$, Genes: $x_{min} = 25$) for which a power law holds. This might reflect that only diseases and genes that are very complex and where a lot of important mechanisms are still unknown, follow the power law. A further observation is that large values of x deviate stronger from the power law fit. We hypothesize that the deviation for large values of x originates from constraints that limit the number of edges, once an entity or a fact is well-known enough. Such a constraint could be the age of a fact. When a fact is known long enough and if it has already been subject of intense discussion in the literature, than it will be not so likely that further studies will be published about this particular fact. See [3] for scale-free networks with constraints.

As the LHGDN is an evolving network, the final distribution of the investigated empirical quantities is not known. However, the current snapshot already represents interesting insights into the behavior how gene-disease associations are studied in the literature.

LHGDN-specific Genes

Genes that are only found in the LHGDN are of particular interest, as they potentially represent hidden knowledge, since they cannot be found in curated databases. Intuitively, we would expect that LHGDN-specific genes are not discussed very often. The probability that a researcher who searches for information about a specific disease stumbles across this particular piece of knowledge is small, since it is not stored in any of the curated databases. Figure 4.7 (a) shows the fraction of LHGDN specific genes with regards to the number of gene mentions. The blue part of the bars represent the fraction of genes that are solely found in the LHGDN for a given number of gene mentions. The main objective of Figure 4.7 (a) is to show the following trend: The probability of finding a LHGDN-specific gene that is part of many facts is rather small. E. g. if a gene is mentioned in the LHGDN exactly once, than the probability of the gene to be LHGDN-specific is approx. 72%. One reason might be that LHGDN-specific genes in facts are quite young and not stored in databases up to now. Figure 4.7 (b) plots the CDF of the LHGDN-specific genes, constituting that



(a)

(b)

Figure 4.7: **Properties of LHGDN-specific genes. (a)** Relative frequency of LHGDN-specific genes vs. number of gene mentions. **(b)** CDF of LHGDN-specific genes.

most of the probability mass is located for small values of gene mentions. Approx. 80% of the probability mass can be found for gene mentions that are mentioned less than five times.

4.5 Conclusions

Summary *Text2SemRel* was applied to the complete GeneRIF database in order to extract semantic gene-disease associations. Hereby, the semantics of the associations constitute a wide variety of biomolecular conditions. To the best of our knowledge, even though the GeneRIF database is relatively small in size, the resulting database, the literature derived human gene-disease network (LHGDN), is currently the largest gene-disease repository publicly available[17]. Furthermore, the LHGDN is integral part of the Linked Life Data[18] initiative.

The LHGDN was compared to several curated state of the art databases (OMIM, UniProt, PharmGKB, CTD), and a union of these (CURATED). CURATED provides an integrated view over the single databases and was constructed with the help of a gene-disease association ontology created by biomedical experts [83]. The number of potential disease genes in the LHGDN is exceeding the number of disease genes that are currently stored in curated databases by far. In total, the LHGDN includes 6.154 potential disease genes. The largest number of disease genes we found on the web, was 4.661 from the GeneCards[19] database (as of April 20th, 2010). GeneCards classifies disease genes according to whether genes cause, predispose to diseases or protect from diseases. GeneCards cannot be downloaded.

All investigated data repositories were analyzed with respect to disease gene properties. So far, such an analysis has only been conducted for the most conservative database, OMIM. The homogeneity analysis shows that genes associated to a disease tend to share similar functional annotations (in terms of pathway annotations and GO annotations). For all investigated databases, the trend to share similar functional annotations differs heavily from random control trials. As diseases become more complex and the number of genes per disease increases, homogeneity drops significantly.

Last but not least, we investigated how entity and fact mentions are distributed in the LHGDN. A careful statistical analysis of the current snapshot of the LHGDN reveals that the number of entity and fact mentions follows a power law distribution.

Discussion Compared to other state of the art databases, one of the unique features in the LHGDN is the classification of gene-disease associations into a broad range of biomolecular conditions. We believe that the chosen characterization of associations represents a valuable contribution for a better understanding of the underlying mechanisms of diseases. In addition, the chosen classification scheme can aid researchers in future experiment design. Therefore, it is crucial for the knowledge discovery process to put extracted facts into the right context by classifying them according to their actual meaning. However, other approaches focus on different but not less important classifications schemes (see e. g. [158]). In many cases a unified view will be needed and data integration issues will arise. Even though the facts stored in the LHGDN are of high quality, they do not represent manually curated knowledge. Future improvements will involve a more advanced entity normalization strategy.

As seen in this chapter, the investigated databases use different controlled vocabularies that vary in granularity. This makes accurate data integration challenging. Given that all, or at least most of the available knowledge is somewhere available in textual form and thus can, in principle be extracted, text mining approaches have an appealing property: they can

[17] `http://www.dbs.ifi.lmu.de/~bundschu/LHGDN.html`

[18] `http://linkedlifedata.com/sources`

[19] `http://www.genecards.org/index.shtml`

alleviate the data integration problem by defining consistent and controlled vocabularies in advance and search subsequently through all available unstructured data sources.

A particular interesting direction for future research is to study the ranking of LHGDN-specific genes with the help of existing disease gene prioritization tools (see [112] for an overview of existing tools). This will help researchers to filter out false positives. Concluding, we hope that the LHGDN will be a valuable knowledge discovery asset for biomedical researchers.

Chapter 5

Summary and Outlook

Summary This thesis has presented advanced techniques for extracting information from unstructured textual resources. All contributions in this thesis are based on the powerful framework of Probabilistic Graphical Models, which can deal with essential problems occurring in almost every real-world application: uncertainty and complexity.

With *Text2SemRel* we introduced a framework that is able to extract facts from textual resources and thus contributes greatly to reduce the gap between text and knowledge. *Text2SemRel* allows expressive search and interactive exploration over the extracted knowledge base and thus facilitates knowledge discovery tremendously. As a result from applying *Text2SemRel* to a large biomedical text collection, we presented the LHGDN, which is currently the largest gene-disease association repository publicly available. The LHGDN is integral part of the Linked Life Data initiative, which confirms the high quality of facts extracted with our presented system.

The TC and the UTC model represent two probabilistic topic models that simulate the generative process of generating semantically annotated document collections. The models can answer a broad range of probabilistic queries and feature flexible capabilities with direct impact on several application fields such as information retrieval, concept search, prediction of annotations, query expansion, probabilistic browsing, and computing document similarities.

Outlook Possible extensions of the presented ideas and contributions in this thesis have been discussed at the end of each of the single chapters.

Broadly speaking, several important properties concerning knowledge in general are covered only very sparsely in current state of the art systems. Taking these inherent properties of knowledge into account, outline major future research directions. For instance, the temporal aspect of knowledge is considered only in a tiny fraction of knowledge extraction systems. Many facts are valid only for a limited space of time. Let's assume that a drug has been approved for the treatment of a disease, but unexpected side effects appear and the Food and Drug Administration cancels the approval. As another example, companies get acquired by other companies and thus disappear. Also CEOs change companies quite often. Considering this limited validity of facts remains a major challenge for information extraction. Another future direction for research concerns the fact that knowledge is encoded in multiple languages. The major fraction of knowledge extraction systems and large-scale knowledge bases are biased towards the English language, but it is disputable if the English language will always lead the way of the future internet. Finally, knowledge extracted from the literature exhibits uncertainty. Besides the inclusion of false positive facts, some facts are highly controversial and might be also subjective. Querying uncertain knowledge will implicate the need for new flexible ranking models that are able to capture highly heterogeneous information needs.

Appendix A

Supporting Information Chapter 4

A.1 Additional Figures

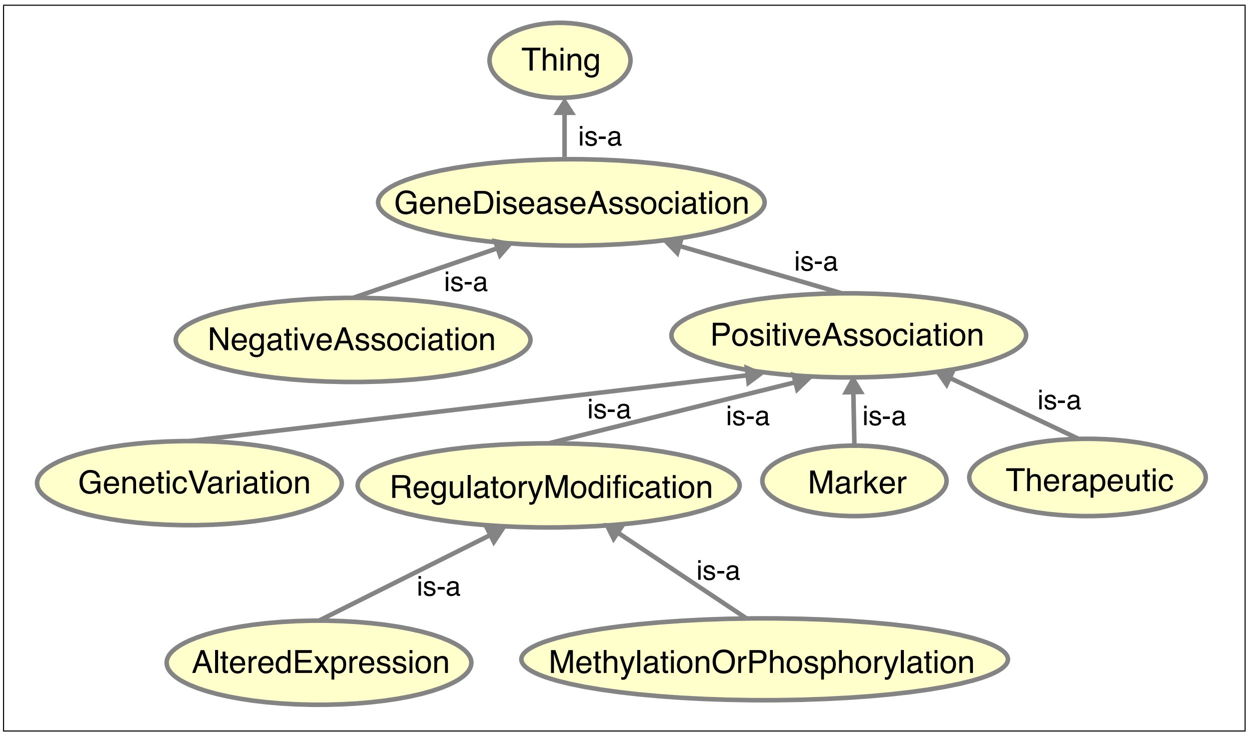
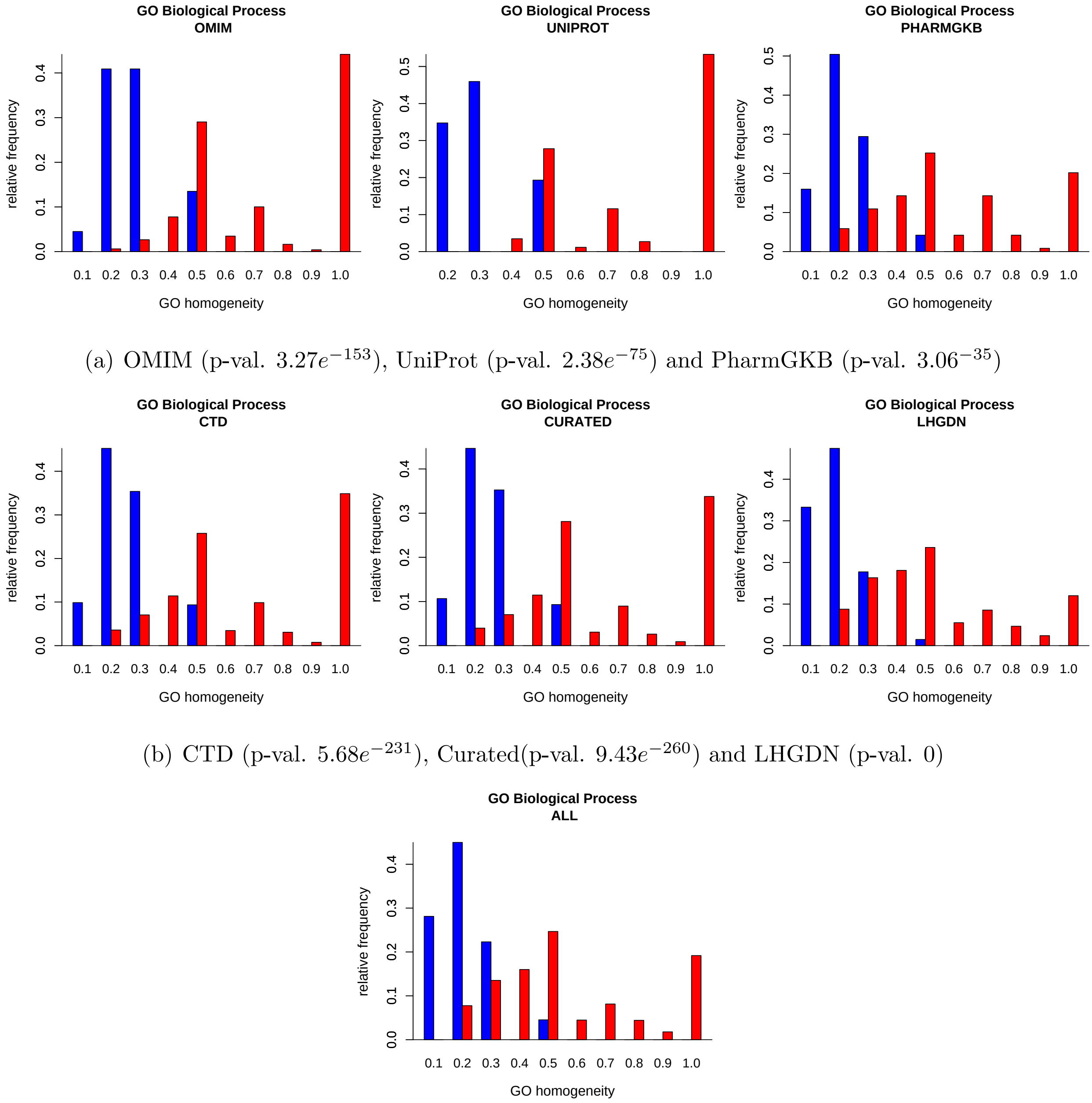


Figure A.1: **Gene-disease association ontology.** The ontology was used to integrate the different databases from Section 4.2. The LHGDN covers *MethylationOrPhosphorylation*, *AlteredExpression*, *GeneticVariation*, *PositiveAssociation*, and *NegativeAssociation*.

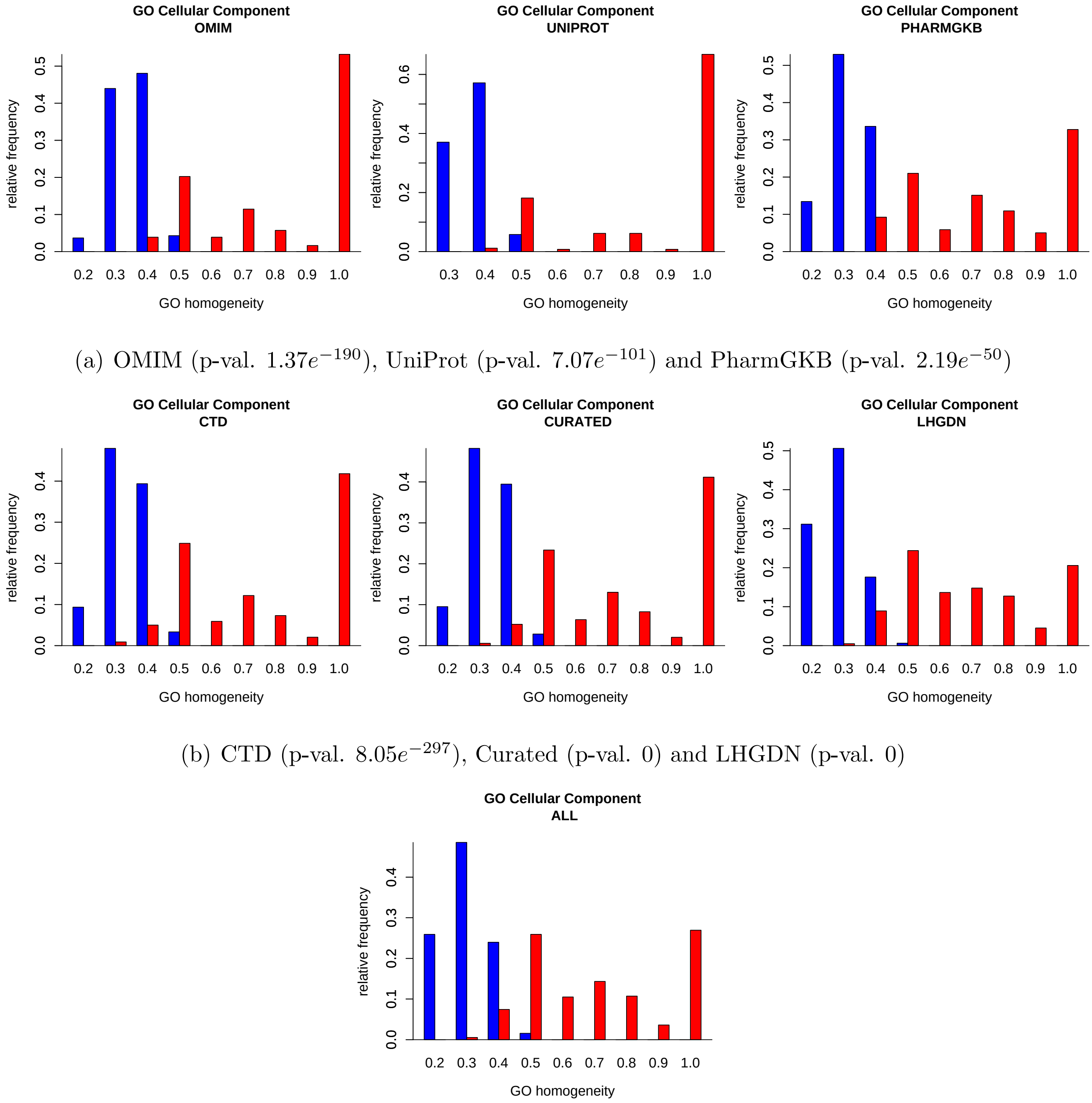


(a) OMIM (p-val. $3.27e^{-153}$), UniProt (p-val. $2.38e^{-75}$) and PharmGKB (p-val. 3.06^{-35})

(b) CTD (p-val. $5.68e^{-231}$), Curated(p-val. $9.43e^{-260}$) and LHGDN (p-val. 0)

(c) ALL (p-val. 0)

Figure A.2: **GO homogeneity (biological process) of the different databases.** The p-values indicate the probability that the two distributions (random and observed) are the same. The x-axis represents the homogeneity values and the y-axis shows the fraction of homogeneity values obtained.

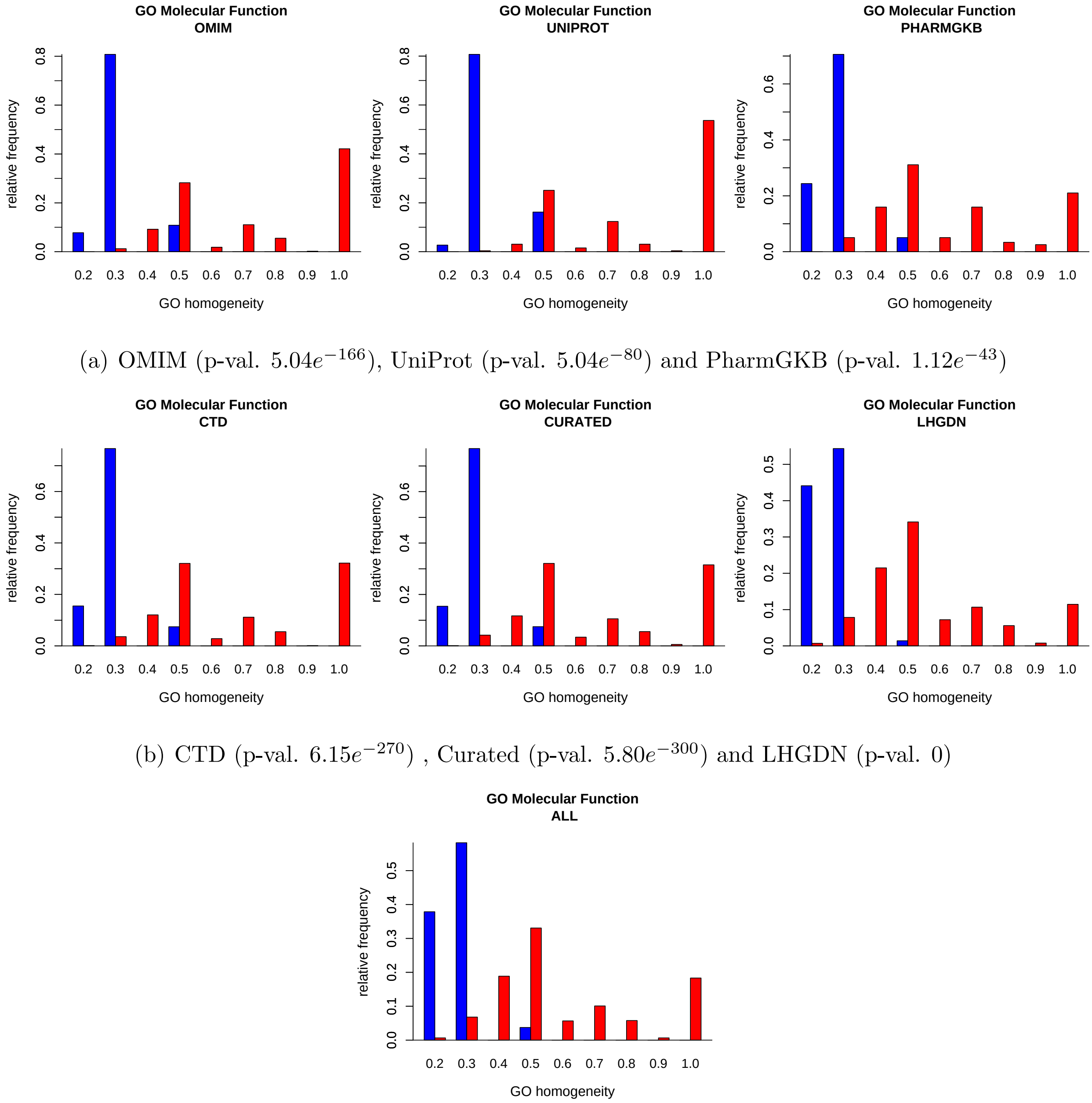


(a) OMIM (p-val. $1.37e^{-190}$), UniProt (p-val. $7.07e^{-101}$) and PharmGKB (p-val. $2.19e^{-50}$)

(b) CTD (p-val. $8.05e^{-297}$), Curated (p-val. 0) and LHGDN (p-val. 0)

(c) ALL (p-val. 0)

Figure A.3: **GO homogeneity (cellular component) of the different databases.** The p-values indicate the probability that the two distributions (random and observed) are the same. The x-axis represents the homogeneity values and the y-axis shows the fraction of homogeneity values obtained.

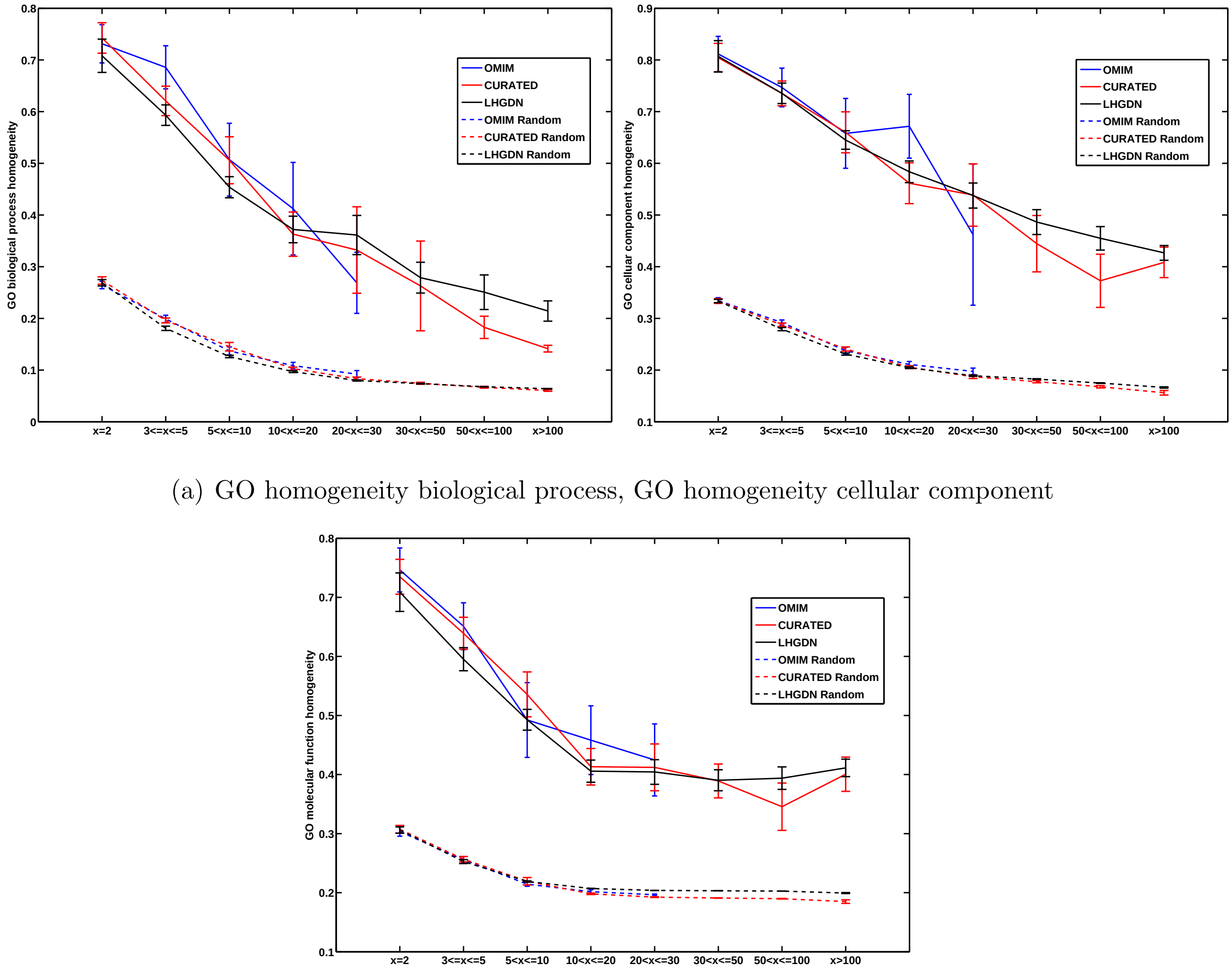


(a) OMIM (p-val. $5.04e^{-166}$), UniProt (p-val. $5.04e^{-80}$) and PharmGKB (p-val. $1.12e^{-43}$)

(b) CTD (p-val. $6.15e^{-270}$) , Curated (p-val. $5.80e^{-300}$) and LHGDN (p-val. 0)

(c) ALL (p-val. 0)

Figure A.4: **GO homogeneity (molecular function) of the different databases.** The p-values indicate the probability that the two distributions (random and observed) are the same. The x-axis represents the homogeneity values and the y-axis shows the fraction of homogeneity values obtained.



(a) GO homogeneity biological process, GO homogeneity cellular component

(b) GO homogeneity molecular function

Figure A.5: **GO homogeneity plots that show the development of homogeneity with respect to the number of genes per disease.** Error bars represent the standard error of the mean (95% confidence interval). The x-axis represents the number of genes per disease.

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

ACE	= Automatic Content Extraction.
CCDF	= Complementary Cumulative Distribution Function.
CDF	= Cumulative Distribution Function.
CoNLL	= Conference on computational Natural Language Learning.
COPD	= Chronic Obstructive Pulmonary Disease.
CRF	= Conditional Random Field.
CTD	= Comparative Toxicogenomics Database.
CVD	= Cardiovascular Disease.
DIPRE	= Dual Iterative Pattern Relation Extraction.
EPO	= Erythropoietin.
ER graph	= Entitiy-Relationship graph.
FN	= False Negative.
FP	= False Positive.
GeneRIF	= Gene Reference Into Function.
GM	= Graphical Model.
GO	= Gene Ontology.
HMM	= Hidden Markov Model.
IE	= Information Extraction.
KDT	= Knowledge Discovery in Textual Databases.
KL	= Kullback-Leibler.
KS	= Kolmogorov-Smirnov.
LD	= Linked Data.
LDA	= Latent Dirichlet Allocation.
LHGDN	= Literature-derived Human Gene-Disease Network.
LLD	= Linked Life Data.
LSI	= Latent Semantic Indexing.
MCMC	= Markov Chain Monte Carlo.
MeSH	= Medical Subject Heading.
MLE	= Maximum Likelihood Estimation.
MRF	= Markov Random Field.
MUC	= Message Understanding Conference.
NB	= Naive Bayes.

NDCG	= Normalized Discounted Cumulative Gain.
NER	= Named Entity Recognition.
NLM	= National Library of Medicine.
NLP	= Natural Language Processing.
NN	= Neural Network.
NP	= Noun Phrase.
OMIM	= Online Mendelian Inheritance in Man.
PGM	= Probabilistic Graphical Model.
PharmGKB	= Pharmacogenomics Knowledge Base.
PLSI	= Probabilistic Latent Semantic Indexing.
POS	= Part-Of-Speech.
RDBMS	= Relational Database Managment System.
RDF	= Resource Description Framework.
RE	= Relation Extraction.
SRE	= Semantic Relation Extraction.
SVD	= Singular Value Decomposition.
SVM	= Support Vector Machine.
TC model	= Topic-Concept model.
TP	= True Positive.
UMLS	= Unified Medical Language System.
UniProt	= Universal Protein Resource.
URI	= Uniform Resource Identifier.
URL	= Uniform Resource Locator.
UTC model	= User-Topic-Concept model.
WWW	= World Wide Web.
YAGO	= Yet Another Great Ontology.

Bibliography

[1] E. Agichtein and L. Gravano. Snowball: extracting relations from large plain-text collections. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94, New York, NY, USA, 2000. ACM.

[2] S. Agrawal, S. Chaudhuri, and G. Das. Dbxplorer: A system for keyword-based search over relational databases. In *ICDE '02: Proceedings of the 18th International Conference on Data Engineering*, pages 5+, Washington, DC, USA, 2002. IEEE Computer Society.

[3] R. Z. Albert and A.-l. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 2001.

[4] W. L. Allen and P. G. Johnston. Have we made progress in pharmacogenomics? the implementation of molecular markers in colon cancer. *Pharmacogenomics*, 6(6):603–614, September 2005.

[5] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. Uniprot: the universal protein knowledgebase. *Nucl. Acids Res.*, 32(suppl_1):D115–119, January 2004.

[6] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA Annual Symposium Proceedings*, pages 17–21, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. alan@nlm.nih.gov, 2001.

[7] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers. The nlm indexing initiative's medical text indexer. In *Medinfo 2004*, pages 268–272. IOS Press, 2004.

[8] M. Ashkenazi, G. D. Bader, A. Kuchinsky, M. Moshelion, and D. J. States. Cytoscape esp: simple search of complex biological networks. *Bioinformatics (Oxford, England)*, 24(12):1465–1466, June 2008.

[9] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[10] A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock, and S. L. Zeger. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther*, 69(3):89–95, March 2001.

[11] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, editors, *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, volume 4825, chapter 52, pages 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg, November 2007.

[12] L. Azzopardi, M. Girolami, and K. van Risjbergen. Investigating the relationship between language model perplexity and ir precision-recall measures. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 369–370, New York, NY, USA, 2003. ACM.

[13] R. Baeza-Yates and P. Raghavan. Chapter 2: Next generation web search. In S. Ceri and M. Brambilla, editors, *Search Computing*, volume 5950, chapter 2, pages 11–23. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[14] M. Banko, M. J. Cafarella, S. Soderl, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI))*, pages 2670–2676, 2007.

[15] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[16] A. Bauer-Mehren, L. I. Furlong, M. Rautschka, and F. Sanz. From snps to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC bioinformatics*, 10 Suppl 8(Suppl 8):S6+, 2009.

[17] F. Belleau, M.-A. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716, October 2008.

[18] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. Swoosh: A generic approach to entity resolution. *VLDB Journal*, 2008.

[19] G. Bhalotia, C. Nakhe, A. Hulgeri, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using banks. In *ICDE*, 2002.

[20] C. Biemann. Ontology learning from text: A survey of methods. *LDV-Forum*, 20(2):75–93, 2005.

[21] A.-L. L. Bienvenu, J. Ferrandiz, K. Kaiser, C. Latour, and S. Picot. Artesunate-erythropoietin combination for murine cerebral malaria treatment. *Acta tropica*, February 2008.

[22] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007.

[23] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[24] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, pages 60–67, 1999.

[25] D. Blei and J. McAuliffe. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, Cambridge, MA, 2008. MIT Press.

[26] D. M. Blei, K. Franks, M. I. Jordan, and I. S. Mian. Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span. *BMC Bioinformatics*, 7(1), May 2006.

[27] D. M. Blei, M. I. Jordan, J. Callan, G. Cormack, C. Clarke, D. Hawking, and A. Smeaton. Modeling annotated data. *SIGIR Forum*, (SPEC. ISS.):127–134, 2003.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[29] S. Blohm and P. Cimiano. Harvesting relations from the web - quantifying the impact of filtering functions. In *Proceedings of the 22nd Conference on Artificial Intelliegence (AAAI'07)*, pages 1316–1323, 2007.

[30] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucl. Acids Res.*, 32(suppl_1):D267–270, January 2004.

[31] B. Bollobas. *Random Graphs*. Cambridge University Press, January 2001.

[32] S. Brin. Extracting patterns and relations from the world wide web. In *WebDB '98: Selected papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, London, UK, 1999. Springer-Verlag.

[33] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[34] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H. P. Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9(1):207+, 2008.

[35] M. Bundschus, M. Dejori, S. Yu, V. Tresp, and H. P. Kriegel. Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text. In *Proceedings of the 8th International Workshop on Data Mining in Bioinformatics (BIOKDD '08)*, 2008.

[36] M. Bundschus, V. Tresp, and H.-p. Kriegel. Topic models for semantically annotated document collections. In *NIPS workshop: Applications for Topic Models: Text and Beyond*, 2009.

[37] M. Bundschus, S. Yu, V. Tresp, A. Rettinger, M. Dejori, and H.-P. Kriegel. Hierarchical bayesian models for collaborative tagging systems. In *IEEE International Conference on Data Mining series (ICDM 2009)*, 2009.

[38] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155, February 2005.

[39] R. C. Bunescu and R. J. Mooney. A shortest path dependency kernel for relation extraction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[40] R. C. Bunescu and R. J. Mooney. Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems*, 2005.

[41] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[42] M. E. Califf, R. J. Mooney, and D. Cohn. Relational learning of pattern-match rules for information extraction. In *In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 328–334, 1999.

[43] G. Carvalho, C. Lefaucheur, C. Cherbonnier, D. Metivier, A. Chapel, M. Pallardy, M.-F. Bourgeade, B. Charpentier, F. Hirsch, and G. Kroemer. Chemosensitization by erythropoietin through inhibition of the nf-[kappa]b rescue pathway. *Oncogene*, aop(current), December 2004.

[44] C. Casals-Pascual, R. Idro, N. Gicheru, S. Gwer, B. Kitsao, E. Gitau, R. Mwakesi, D. J. Roberts, and C. R. Newton. High levels of erythropoietin are associated with protection against neurological sequelae in african children with cerebral malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(7):2634–2639, February 2008.

[45] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, 2008.

[46] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines, 2001.

[47] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines:augmenting social networks with text. In *KDD '09: Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[48] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 2009.

[49] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *7th International Semantic Web Conference (ISWC 2008)*, 2008.

[50] H. Chen and B. M. Sharp. Content-rich biological network constructed by mining pubmed abstracts. *BMC bioinformatics*, 5(1):147+, October 2004.

[51] P. P. Chen. The entity-relationship model: Toward a unified view of data. *ACM Transactions on Database Systems*, 1:9–36, 1976.

[52] S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, 1999.

[53] Y.-W. Chen and C.-J. Lin. Combining svms with various feature selection strategies. In I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, chapter 13, pages 315–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[54] H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Automatic recognition of topic-classified relations between prostate cancer and genes using medline abstracts. *BMC Bioinformatics*, 7(Suppl 3):S4+, 2006.

[55] H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 4–15, 2006.

[56] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 462–471, New York, NY, USA, 2004. ACM.

[57] P. Cimiano, G. Ladwig, and S. Staab. Gimme' the context: context-driven automatic semantic annotation with c-pankow. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 332–341, New York, NY, USA, 2005. ACM.

[58] P. Cimiano and J. Valker. *Text2Onto: A Framework for Ontology Learning and Data-driven Change Discovery.* 2005.

[59] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661+, Feb 2009.

[60] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366–2382, September 2007.

[61] K. Cohen and L. Hunter. Natural language processing and systems biology. In W. Dubitzky and F. Azuaje, editors, *Artificial Intelligence Methods And Tools For Systems Biology*, volume 5 of *Computational Biology*, chapter 9, pages 147–173. Springer Netherlands, Dordrecht, 2004.

[62] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky. Emergent behavior of growing knowledge about molecular interactions. *Nature Biotechnology*, 23(10):1243–1247, October 2005.

[63] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 1999.

[64] A. Culotta, A. Mccallum, and J. Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Human Language Technology Conference/North American chapter of the Association for Computational Linguistics Annual Meeting*, June 2006.

[65] N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A web of concepts. In *PODS '09: Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–12, New York, NY, USA, 2009. ACM.

[66] M. G. de Verdier. The big three concept: a way to tackle the health care crisis? *Proceedings of the American Thoracic Society*, 5(8):800–805, December 2008.

[67] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[68] T. M. Department, T. Minka, and J. Lafferty. Expectation-propagation for the generative aspect model. In *In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[69] T. G. Dietterich. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, London, UK, 2002. Springer-Verlag.

[70] T. G. Dietterich, A. Ashenfelter, and Y. Bulatov. Training conditional random fields via gradient tree boosting. In *In Proceedings of the 21th International Conference on Machine Learning (ICML)*, pages 217–224, 2004.

[71] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *Proceedings of LREC 2004*, pages 837–840, 2004.

[72] P. Drucker. *The Age of Discontinuity: Guidelines to Our Changing Society.* Transaction Publishers, January 1992.

[73] P. F. Drucker. Knowledge-worker productivity: The biggest challenge. *California Management Review*, 41:79–41, 1999.

[74] F. Erik, T. K. Sang, and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In W. Daelemans and M. Osborne, editors, *Proceedings of the Conference on Natural Language Learning (CoNLL-2003)*, pages 142–147, Edmonton, Kanada, 2003. Association for Computational Linguistics (ACL).

[75] E. Erosheva, S. Fienberg, and J. La. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 2004.

[76] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December 2008.

[77] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134, June 2005.

[78] I. Feldman, A. Rzhetsky, and D. Vitkup. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11):4323–4328, March 2008.

[79] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95*, pages 112–117, 1995.

[80] C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, illustrated edition edition, May 1998.

[81] P. Flaherty, G. Giaever, J. Kumm, M. I. Jordan, and A. P. Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics (Oxford, England)*, 21(15):3286–3293, August 2005.

[82] K. Fundel, R. Kueffner, and R. Zimmer. Relex - relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[83] L. Furlong, A. Bauer-Mehren, M. Bundschus, and F. Sanz. Network analysis of a comprehensive database of gene-disease associations. *Manuscript in preparation for submission to Molecular Systems Biology*, 2010.

[84] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[85] Y. Garten and R. B. Altman. Teaching computers to read the pharmacogenomics literature ... so you don't have to. *Pharmacogenomics*, 11(4):515–518, April 2010.

[86] C. W. Gay, M. Kayaalp, and A. R. Aronson. Semi-automatic indexing of full text biomedical articles. In *AMIA Annu Symp Proc*, pages 271–275, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD 20894, USA., 2005.

[87] C. B. Giles and J. D. Wren. Large-scale directional relationship extraction and resolution. *BMC bioinformatics*, 9 Suppl 9(Suppl 9):S11+, 2008.

[88] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, May 2007.

[89] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April 2004.

[90] R. Grishman and B. Sundheim. Design of the muc-6 evaluation. In *Proceedings of the 6th conference on Message Understanding*, November 1995.

[91] R. Gupta and S. Sarawagi. Creating probabilistic databases from information extraction models. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 965–976. VLDB Endowment, 2006.

[92] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(Database issue):D514–517, January 2005.

[93] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, de La, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The gene ontology (go) database and informatics resource. *Nucl. Acids Res.*, 32(suppl_1):D258–261, January 2004.

[94] M. Harvey, M. Baillie, I. Ruthven, and M. Carman. Tripartite hidden topic models for personalised tag suggestion. In *Advances in Information Retrieval: 32nd European Conference on IR Research*, 2010.

[95] M. A. Hearst. *Information Visualization for Text Analysis.* Cambridge University Press, 1 edition, September 2009.

[96] G. Heinrich. Parameter estimation for text analysis,. Technical report, University of Leipzig, 2008.

[97] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein. Pharmgkb: the pharmacogenetics knowledge base. *Nucleic acids research*, 30(1):163–165, January 2002.

[98] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In S. H. Myaeng, D. W. Oard, F. Sebastiani, T. S. Chua, M. K. Leong, S. H. Myaeng, D. W. Oard, F. Sebastiani, T. S. Chua, and M. K. Leong, editors, *Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[99] L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. Overview of biocreative task 1b: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1, 2005.

[100] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1, 2005.

[101] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[102] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. *Information Retrieval in Folksonomies: Search and Ranking*, volume 4011 of *Lecture Notes in Computer Science*, chapter 31, pages 411–426. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[103] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, 74(2-4):289–298, March 2005.

[104] S. Humphrey, C. Lu, W. Rogers, and A. Browne. Journal descriptor indexing tool for categorizing text according to discipline or semantic type. In *AMIA Annu Symp Proc*, 2006.

[105] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. C. Rindflesch. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. *J Am Soc Inf Sci Technol*, 57(1):96–113, 2006.

[106] I. Iossifov, R. Rodriguez-Esteban, I. Mayzus, K. J. Millen, and A. Rzhetsky. Looking at cerebellar malformations through text-mined interactomes of mice and humans. *PLoS computational biology*, 5(11):e1000559+, November 2009.

[107] F. C. J. Iria. Relation extraction for mining the semantic web. In *Dagstuhl Seminar on Machine Learning for the Semantic Web*, 2005.

[108] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in Information Retrieval*, 2000.

[109] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, pages 506–514. Springer-Verlag, Berlin, Heidelberg, 2007.

[110] T. S. Jayram, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu. Avatar information extraction system. *IEEE Data Engineering Bulletin*, 2006.

[111] J. Jiang and C. Zhai. A systematic exploration of the feature space for relation extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 113–120, Rochester, New York, April 2007. Association for Computational Linguistics.

[112] M. G. Kann. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinform*, 11(1):96–110, January 2010.

[113] G. Kasneci, M. Ramanath, F. Suchanek, and G. Weikum. The yago-naga approach to knowledge discovery. *SIGMOD Rec.*, 37(4):41–47, 2008.

[114] J. Kazama and K. Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, 2007.

[115] R. Klinger and K. Tomanek. Classical probabilistic models and conditional random fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, December 2007.

[116] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In L. D. Bergman, A. Tuzhilin, R. Burke, A. Felfernig, L. S. Thieme, L. D. Bergman, A. Tuzhilin, R. Burke, A. Felfernig, and L. S. Thieme, editors, *RecSys*, pages 61–68. ACM, 2009.

[117] I. Kunz, M.-C. C. Lin, and L. Frey. Metadata mapping and reuse in cabig. *BMC bioinformatics*, 10 Suppl 2, 2009.

[118] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *22nd Annual Conference on Neural Information Processing Systems*, 2008.

[119] J. Lafferty, A. Mccallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[120] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project.* Addison-Wesley Pub (Sd), April 1990.

[121] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[122] Y. Li and P. Agarwal. A pathway-based view of human diseases and disease relationships. *PloS one*, 4(2):e4346+, February 2009.

[123] H. J. Lowe and G. O. Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *JAMA : the journal of the American Medical Association*, 271(14):1103–1108, April 1994.

[124] Z. Lu, K. B. Cohen, and L. Hunter. Generif quality assurance as summary revision. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 269–280, 2007.

[125] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms.* Cambridge University Press, 1st edition, June 2002.

[126] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 33(Database issue), January 2005.

[127] M. Masseroli, H. Kilicoglu, F.-M. Lang, and T. C. Rindflesch. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC bioinformatics*, 7:291+, June 2006.

[128] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research*, 37(Database issue):D619–622, January 2009.

[129] C. J. Mattingly, M. C. Rosenstein, A. P. P. Davis, G. T. Colby, J. N. Forrest, and J. L. Boyer. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicological sciences : an official journal of the Society of Toxicology*, 92(2):587–595, August 2006.

[130] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *The Fifteenth National Conference on Artificial Intelligence (AAAI*, 1998.

[131] A. Mccallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.

[132] A. K. McCallum. Mallet: A machine learning for language toolkit. 2002.

[133] D. M. McDonald, H. Chen, H. Su, and B. B. Marshall. Extracting gene pathway relations using a hybrid grammar: the arizona relation parser. *Bioinformatics*, 20(18):3370–3378, December 2004.

[134] V. A. McKusick. Mendelian inheritance in man and its online version, omim. *American journal of human genetics*, 80(4):588–604, April 2007.

[135] J. A. Mitchell, A. R. Aronson, J. G. Mork, L. C. Folk, S. M. Humphrey, and J. M. Ward. Gene indexing: characterization and analysis of nlm's generifs. In *AMIA Annu Symp Proc*, pages 460–464, University of Missouri – Columbia, USA., 2003.

[136] T. M. Mitchell. *Machine Learning.* McGraw-Hill Science/Engineering/Math, 1 edition, March 1997.

[137] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1:226–251, 2004.

[138] F. Mörchen, M. Dejori, D. Fradkin, J. Etienne, B. Wachmann, and M. Bundschus. Anticipating annotations and emerging trends in biomedical literature. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 954–962, New York, NY, USA, 2008. ACM.

[139] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. H. Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. Overview of biocreative ii gene normalization. *Genome biology*, 9 Suppl 2(Suppl 2), 2008.

[140] I. Mukherjee and D. M. Blei. Relative performance guarantees for approximate inference in latent dirichlet allocation. In *Neural Information Processing Systems*, 2008.

[141] A. Névéol, S. E. Shooshan, S. M. Humphrey, T. C. Rindflesch, and A. R. Aronson. Multiple approaches to fine-grained indexing of the biomedical literature. In R. B. Altman, K. A. Dunker, L. Hunter, T. Murray, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 292–303. World Scientific, 2007.

[142] A. Névéol, S. E. Shooshan, J. G. Mork, and A. R. Aronson. Fine-grained indexing of the biomedical literature: Mesh subheading attachment for a medline indexing tool. In *Proc. AMIA Symp*, 2007.

[143] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686, New York, NY, USA, 2006. ACM.

[144] Z. Nie, Y. Ma, S. Shi, J. R. Wen, and W. Y. Ma. Web object retrieval. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 81–90, New York, NY, USA, 2007. ACM.

[145] S. D. Nimer. Myelodysplastic syndromes. *Blood*, 111(10):4841–4851, May 2008.

[146] J. Osborne, J. Flatow, M. Holko, S. Lin, W. Kibbe, L. Zhu, M. Danila, G. Feng, and R. Chisholm. Annotating the human genome with disease ontology. *BMC Genomics*, 10(Suppl 1):S6+, 2009.

[147] A. Ozgür, T. Vu, G. Erkan, and D. R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics (Oxford, England)*, 24(13):i277–285, July 2008.

[148] P. Pantel and M. Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[149] Paul, S. Costache, W. Nejdl, and S. Handschuh. P-tag: large scale automatic generation of personalized annotation tags for the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 845–854, New York, NY, USA, 2007. ACM.

[150] M. F. Porter. An algorithm for suffix stripping. *Readings in Information Retrieval*, pages 313–316, 1997.

[151] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1165–1172 vol.2, 1999.

[152] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[153] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[154] D. Ramage, P. Heymann, C. D. Manning, and H. G. Molina. Clustering the tagged web. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63, New York, NY, USA, 2009. ACM.

[155] A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):R40+, 2005.

[156] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In D. Yarovsky and K. Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey, 1995. Association for Computational Linguistics.

[157] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[158] T. C. Rindflesch, B. Libbus, D. Hristovski, A. R. Aronson, and H. Kilicoglu. Semantic relations asserting the etiology of genetic diseases. In *AMIA Annu Symp Proc*, pages 554–558, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20894, USA., 2003.

[159] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of Pacific Symposium on Biocomputing*, pages 517–528, 2000.

[160] P. M. Roberts and W. S. Hayes. Information needs and the role of text mining in drug development. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 592–603, 2008.

[161] R. Rodriguez-Esteban, I. Iossifov, and A. Rzhetsky. Imitating manual curation of text-mined facts in biomedicine. *PLoS Comput Biol*, 2(9):e118+, September 2006.

[162] B. Rosario and A. Hearst. Multi-way relation classification: Application to protein-protein interaction. In *Human Language Technology Conference on Empirical Methods in Natural Language Processing*, 2005.

[163] B. Rosario and M. Hearst. Classifying semantic relations in bioscience texts. In *Proceedings of the Annual Meeting of Association of Computational Linguistics (ACL '04)*, 2004.

[164] R. Rubinstein and I. Simon. Milano–custom annotation of microarray results using automatic literature searches. *BMC bioinformatics*, 6(1), 2005.

[165] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboué, W. Weng, W. J. Wilbur, V. Hatzivassiloglou, and C. Friedman. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of biomedical informatics*, 37(1):43–53, February 2004.

[166] A. Rzhetsky, M. Seringhaus, and M. Gerstein. Seeking a new biology through text mining. *Cell*, 134(1):9–13, July 2008.

[167] A. Rzhetsky, H. Shatkay, and J. J. Wilbur. How to get the most out of your curation effort. *PLoS Comput Biol*, 5(5):e1000391+, May 2009.

[168] S. Sarawagi. Information extraction. *Found. Trends databases*, 1(3):261–377, 2008.

[169] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In *Data Science and Classification*, pages 261–270. 2006.

[170] S. R. Seong, J. W. Lee, Y. K. Lee, T. I. Kim, D. J. Son, D. C. Moon, Y. W. Yun, d. o. . Y. Yoon, and J. T. Hong. Stimulation of cell growth by erythropoietin in raw264.7 cells: association with ap-1 activation. *Archives of pharmacal research*, 29(3):218–223, March 2006.

[171] W. Shen, A. Doan, J. F. Naughton, and R. Ramakrishnan. Declarative information extraction using datalog with embedded extraction predicates. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 1033–1044. VLDB Endowment, 2007.

[172] T. Shtatland, D. Guettler, M. Kossodo, M. Pivovarov, and R. Weissleder. Pepbank–a database of peptides based on sequence text mining and public peptide data sources. *BMC bioinformatics*, 8(1):280+, August 2007.

[173] S. Soderland, C. Cardie, and R. Mooney. Learning information extraction rules for semi-structured and free text. In *Machine Learning*, volume 34, pages 233–272, 1999.

[174] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. Mc Namara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. 2007.

[175] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA, 2004. ACM.

[176] F. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 2008.

[177] F. M. Suchanek, G. Ifrim, and G. Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 712–717, New York, NY, USA, 2006. ACM Press.

[178] F. M. Suchanek, M. Sozio, and G. Weikum. Sofie: a self-organizing framework for information extraction. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 631–640, New York, NY, USA, 2009. ACM.

[179] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[180] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *J. Mach. Learn. Res.*, 8:693–723, 2007.

[181] L. Tanabe and W. J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, August 2002.

[182] T. Tatusova. Genomic databases and resources at the national center for biotechnology information. *Methods in molecular biology (Clifton, N.J.)*, 609:17–44, 2010.

[183] T.-H. H. Tsai, S.-H. H. Wu, W.-C. C. Chou, Y.-C. C. Lin, D. He, J. Hsiang, T.-Y. Y. Sung, and W.-L. L. Hsu. Various criteria in the evaluation of biomedical named entity recognition. *BMC bioinformatics*, 7(1):92+, February 2006.

[184] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28, January 2006.

[185] A. Vailaya, P. Bluvas, R. Kincaid, A. Kuchinsky, M. Creech, and A. Adler. An architecture for biological information extraction and representation. *Bioinformatics (Oxford, England)*, 21(4):430–438, February 2005.

[186] N. N. Vishwanathan. Accelerated training of conditional random fields with stochastic gradient methods. In *Proceedings of the 23rd international conference on Machine learning*, pages 969–976, 2006.

[187] L. Wiese, C. Hempel, M. Penkowa, N. Kirkby, and J. A. L. Kurtzhals. Recombinant human erythropoietin increases survival and reduces neuronal apoptosis in a murine model of cerebral malaria. *Malaria Journal*, 7:3+, January 2008.

[188] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50, New York, NY, USA, 2007. ACM.

[189] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *WWW2006: Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.

[190] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

[191] L. Yao, D. Mimno, and A. Mccallum. Efficient methods for topic model inference on streaming document collections. In *The 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining (KDD 2009)*, 2009.

[192] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. Biocreative task 1a: gene mention finding evaluation. *BMC Bioinformatics*, 6, 2005.

[193] Y. T. Yen, B. Chen, H. W. Chiu, Y. C. Lee, Y. C. Li, and C. Y. Hsu. Developing an nlp and ir-based algorithm for analyzing gene-disease relationships. *Methods Inf Med*, 45(3):321–329, 2006.

[194] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106, 2003.

[195] S. Zhao and R. Grishman. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 419–426, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[196] B. Zheng, D. C. Mclean, and X. Lu. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, 7, 2006.

[197] G. Zhou, D. Shen, J. Zhang, J. Su, and S. Tan. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC bioinformatics*, 6 Suppl 1, 2005.

[198] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J. R. Wen. Statsnowball: a statistical approach to extracting entity relationships. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 101–110, New York, NY, USA, 2009. ACM.