Saarland University
Faculty of Natural Sciences and Technology I
Department of Computer Science

Bachelor's Thesis

Automatic Extraction and Enrichment of a Multilingual Domain Model

Author:
Kyrill Pugatschewski

submitted on
September 22, 2015

Supervisor:
Prof. Dr. Jörg Siekmann

Advisor:
Dr. Sergey Sosnovsky

Reviewers:
Prof. Dr. Jörg Siekmann
Dr. Sergey Sosnovsky

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken,_____ _____
(Datum/Date) (Unterschrift/Signature)

Acknowledgments

First of all, I would like to express my gratitude to my advisor Dr. Sergey Sosnovsky for his guidance throughout the work on this thesis. Our numerous meetings and discussions (often far beyond the scope of the thesis itself) were a fruitful source of ideas. In particular during later stages of the work, his extensive advice and frank criticism were invaluable to the completion of this document.

Moreover, I would like to thank Prof. Dr. Jörg Siekmann for supervising my thesis, as well as his words of advice on issues more far-reaching than just the current work.

Lastly, a big thank you to Sabrina, not only for proofreading, but for her patience and unconditional support, for just being there, even when my thoughts kept crawling back to work.

Abstract

This thesis presents a framework for automated learning of a semantic, multilingual domain model. It tackles the problem of Ontology Learning by combining automatic extraction of expert knowledge from an online glossary with Linked Open Data. The obtained model contains lexical and semantic information on the domain of probability theory and mathematical statistics in three languages. During the linking of its constituent concepts to DBpedia, a linked open dataset, the problem of ambiguity arises. For one query term, several candidate concepts may be suggested. A multi-pass algorithm is presented which combines different techniques to disambiguate the word sense. Several approaches have been evaluated to come up with the best possible assembly of methods enabling the best precision and recall of the word sense disambiguation process.

Contents

1 Introduction

Domain models are a common backbone for various intelligent services, which rely on them for formal reasoning in terms of elementary knowledge components constituting the domain. One method to create such domain models is manual knowledge engineering, however this is a considerably hard task. It requires profound insight and expertise in the target domain. Even the compilation of domain information by groups of experts is a time-consuming process. Fortunately, there are freely available knowledge sources online that have been created by communities of experts, such as encyclopedias, glossaries and thesauri. In addition, the emergence of the Semantic Web and Linked Open Data lead to a plethora of semantically-linked information likewise freely available online.

The combination of both types of knowledge sources affords an opportunity. For many concepts established in expert resources, one can find corresponding entities in open linked datasets. These can provide additional relations and information beyond the scope of manual compilation, thus allowing the enrichment of expert knowledge. As a result, one can automate the creation of ontologies as comprehensive domain models by means of enriching expert knowledge with Linked Open Data.

1.1 Context of This Work

The work described in this thesis was carried out in the framework of the Interlingua project. Its goal was to develop a web-based system providing individual support to students studying in a foreign language. [1] The project was funded by INTERREG IV-A-GR. For implementation and evaluation, probability theory and mathematical statistics have been chosen as the target subject.

Major efforts have been made in the past few years to turn the European Union into a homogeneous educational area. Initiatives like the Bologna process and Erasmus program have helped to harmonize transnational study schemes and develop the European competence concept. In the Greater Region of Saarland, Rhineland-Palatinate, Lorraine, Wallonia and Luxembourg, the INTERREG program has supported development of language related technology for learning and promoted multilingualism of the youth. Those students face substantial difficulties: the focus of a foreign university's syllabus may follow different educational practices and may require background knowledge not covered at the home university. The background knowledge of the student exists in the native language, while the terminology required in the course is in a foreign language. On top of that, getting accustomed to studying in a foreign language is a challenge itself.

As a use case, one can consider a German student enrolled in the bilingual ISFATES/DFHI program. During their studies in France, they can read French textbooks on the Interlingua platform. When coming across terminology, which is difficult to understand for them, they can get a related reading in German. This related reading is not a direct translation, but comes from a textbook established in the German didactic community and therefore accounts for the teaching practices the student is used to.

[1] http://www.interlingua-project.eu/

The Interlingua service supports the domain of probability theory and mathematical statistics and the three languages English, German and French, allowing students to switch between them while reading educational material. Furthermore they can evaluate their progress by querying self-assessment tests generated from certain keywords in the texts. The project approaches these issues by methods borrowed from the fields of Semantic Web, Natural Language Processing and Information Extraction. Semantic linking between textbooks in different languages provides students with related reading in their mother tongue. To achieve such a semantic linking, a comprehensive, multilingual ontology needs to be used as a reference model helping to establish links between texts with similar meaning in different languages.

1.2 Approach & Contributions

This thesis presents a framework for the automated creation of a multilingual domain model and its enrichment with knowledge from DBpedia. Figure 1 graphically represents the different phases of the approach and the exchange of information between them.

It begins with extracting the base for the model from a well-formatted multilingual online glossary. A SKOS representation is achieved by creating a concept for each entry. Then, these concepts are mapped to corresponding resources in DBpedia. From there, the ontology is enriched with definitions, links to corresponding Wikipedia articles, as well as hierarchical and associative relations between resources in different languages.

While querying DBpedia for the mapping, it is important to consider ambiguities in the labels of the concepts, for a single concept one may discover several candidates on DBpedia. Three techniques are utilized for disambiguation of DBpedia resources:

- A graph built from the links between DBpedia resources is used to discard candidate resources not belonging to a dense sub-graph representing the target domain.
- From a domain corpus of documents consisting of English, German and French textbooks, relevant sections for a given term can be retrieved. This textual information is compared with candidate resources' abstracts or Wikipedia articles in terms of cosine similarity.
- Under the assumption that the abstracts and articles of correct resources contain further statistical terms, queries are built from the lexical labels in the model. With these, one can retrieve correct resources by comparing the cosine similarities between the queries and the candidate resources' abstracts or Wikipedia articles.

This thesis has two main contributions:

- A fully automatic framework for learning multilingual semantic models by leveraging Linked Open Data is presented.
- For the inherent task of Word Sense Disambiguation, a multi-layered algorithm is proposed, which combines three different techniques to maximize the performance.

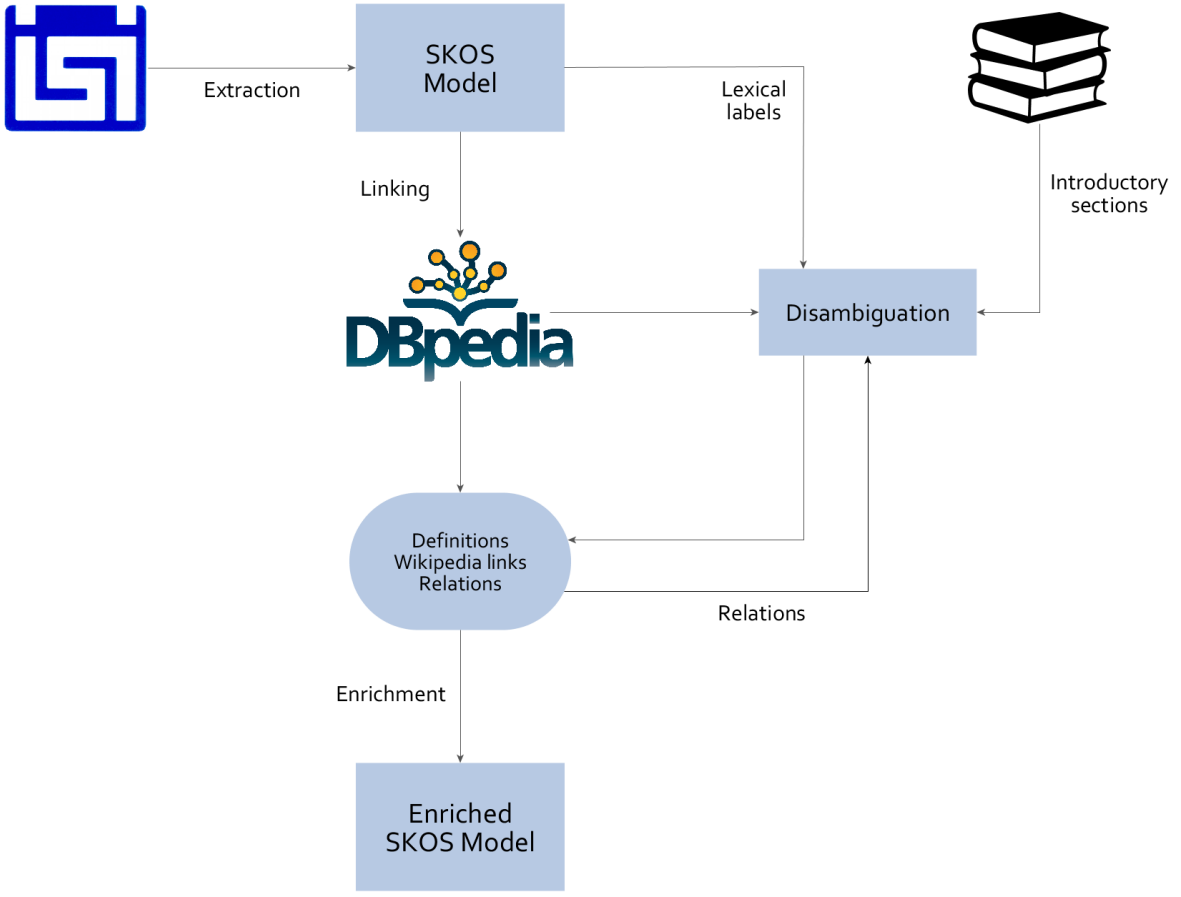


Figure 1: General architecture (`http://www.isi-web.org`, `http://www.dbpedia.org`], Icon made by `http://www.freepik.com` Freepik from `www.flaticon.com` is licensed under `http://creativecommons.org/licenses/by/3.0/` CC BY 3.0. No changes were made.)

1.3 Outline of the Thesis

Structure-wise, the thesis can be divided into four main parts and the conclusion. Chapter 2 introduces the three fields related to this thesis: Ontology Learning, Linked Open Data and Word Sense Disambiguation. In particular, research from the intersection of Ontology Learning and Linked Open Data is presented, which explores the utilization of Linked Open Data as an additional knowledge source during the learning of ontologies.

Chapter 3 deals with the extraction of the model from a well-formatted online glossary. It introduces the ISI Multilingual Glossary of Statistical Terms, which is used as the source of multilingual lexical information for populating the model. An overview of SKOS is given followed by a description of the employed SKOS elements for serializing the model.

The fourth chapter introduces DBpedia as the major open linked dataset used in this thesis. It explains how the model is enriched with semantic information. Additionally, a section approaches the issue of missing links on DBpedia.

A multi-layered algorithm for resolving ambiguity in DBpedia resources is presented in chapter 5. It leverages resource linkage, textual information from a domain corpus and the model itself.

At the end, the results of the presented work are summarized and an outlook on possible future work is given.

2 Background & Related Work

The presented approach follows three stages that use techniques from the three related research fields.

Automated creation of semantic models is researched in the larger field of Ontology Learning, on which general methodology and exemplary work is given. This serves as a background for research more connected to ours, namely the integration of Linked Open Data sources into the ontology learning process. Furthermore, the task of Word Sense Disambiguation is outlined, which has to be addressed during the ontology enrichment. Research on graph-based disambiguation is considered in particular to point out more formal methods in contrast to the rather intuitive approach presented in this thesis. The techniques mentioned in the analyzed papers are also valuable for future work.

2.1 Ontology Learning

Ontology learning is the task of deriving concepts, relations and axioms from various information sources and using them to construct an ontology. The usually employed techniques come from information retrieval, data mining, machine learning as well as natural language processing. Most approaches begin with the extraction of terms from some textual input with respect to their relevance to a domain (termhood) and noun sequences constituting collocations (unithood). By grouping variants of a term, concepts are formed as units of thought. To model the ties between them, two types of relations can be extracted, taxonomic (hierarchical) relations and non-taxonomic relations (e.g. meronymy, attributes, thematic roles). Axioms are learned by discovering logical facts from the input and are used to define further constraints and deduction of other truth.[1]

In the OntoLearn system[2], Navigli et al. perform ontology learning with linguistics and statistics-based techniques. On a part-of-speech tagged domain corpus, they employ the two metrics *domain relevance* and *domain consensus* for term extraction from noun phrases. Domain relevance measures the specificity of a term by normalizing its frequency in the target domain with the frequencies of all other candidate terms in that domain and the frequencies of the term across some general corpora. Domain consensus is an entropic value describing the appearance of a term in a single document as compared to the whole target domain. Identified terms are then compared to existing concepts from WordNet[3] with *semantic interpretation*. It evaluates multi-word terms

by finding intersections of semantic graphs from WordNet, consisting of non-taxonomic relations between synsets. After selecting the best sense combinations by weighting common semantic patters, the non-taxonomic and taxonomic relations are learned by following semantic relations on WordNet.

TextStorm/Clouds[4] by Oliveira et al. applies logic and linguistics-based techniques to perform ontology learning in a semi-automated framework. Instead of plain terms, they extract binary predicates. After pre-processing an input text with part-of-speech tagging, they perform dependency analysis. Using an augmented grammar, they discover relations of two types: relations induced by main verbs between nouns and properties induced by compound nouns. As an exemplary result, they get *eat(Zebra, grass)* and *property(grass, green)* from the sentence *Zebra eat green grass*. Binary predicates are then manually aggregated into hierarchies and other semantic relations into a simple ontology. This ontology is then inspected with inductive logic programming to infer axioms from recurrent concepts and relations in the predicates.

There has also been research on learning ontologies from existing semi-structural knowledge sources. MECUREO[5] has been developed to automatically construct ontologies from dictionaries as domain models for e-learning applications[2]. The approach of Apted and Kay relies on consistent grammatical conventions in dictionary entries. They are captured by a manually defined mapping from keywords to relationships: on an encounter of "as in", they deduce parentship between the current entry and the linked entry; similarly, they derive e.g. antonymy from the keyword "opposed" and a sibling relation from "see".

2.2 Linked Open Data

For the realization of the Semantic Web, two aspects have to be considered. On the one hand, semantic data has to be made available online, on the other hand, the data has to be interlinked to enable exploration. Thus, linked data has been described as being "essential to actually connect the semantic web"[6]. Apart from being machine-readable, four criteria must be met by a data set or an ontology to be considered a part of open linked data: entities must be identified by URIs, URIs must conform to HTTP, information about entities must be provided in RDF standards, there must be links to URIs in other locations. The Linking Open Data initiative[3] has helped to interlink data sets available under open licenses. At the time of writing this thesis, the *Linked Open Data Cloud* consists of 570 data sets[7]. Since the emergence of projects like DBpedia[8] ("a nucleus for a web of open data"[9]), research has been done on employing linked open data for ontology learning.

In [10], Weichselbraun et al. show how integrating external knowledge sources like DBpedia[8] and OpenCyc[11, 12] into ontology learning systems can be used to automatically suggest labels for non-taxonomic relations. A meta-ontology defining a set of

[2]The Free On-Line Dictionary of Computing (http://foldoc.org) has been used for development
[3]http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

relation labels is created by experts which captures the relations in the domain ontology. It also includes domain, range and property restrictions for the relations. Pairs of concepts are than mapped to DBpedia resources to obtain the corresponding formal OpenCyc resource. While traversing the type hierarchy of the OpenCyc resource, they compare the current type with the restrictions from the meta-ontology. If a type is reached which corresponds to some restriction for a relation, this is a good indication for the concept pair to participate in that relation.

Closer to the work presented in this thesis, researchers have investigated the usage of linked open data for semi-automated and automated bootstrapping of thesauri.

PoolParty[13] is a SKOS thesaurus management tool supporting domain experts in the creation and maintenance of thesauri. Schandl et al. identify several cases in which linking thesaurus concepts to DBpedia helps to provide additional metadata. As an exemplary setting, consider a small taxonomy of art. The concept *August Macke* has been previously manually mapped to the corresponding DBpedia resource. PoolParty can now suggest more expressionist painters by following category links on DBpedia for *August Macke* to discover sibling resources as well as DBpedia URIs for the broader concept *Expressionist painters*. This also helps to disambiguate concepts which have not been linked yet. When querying for an expressionist painter, whose name is ambiguous according to DBpedia, it is likely that the candidate resource, which has the corresponding category for *Expressionist painters*, is the correct one. The same approach can be used for linking new unambiguous concepts to DBpedia and assigning a broader concept from the thesaurus. Furthermore, linked concepts can be enriched with linked open data information. Schandl et al. mention the appropriation of DBpedia abstracts as definitions and alternate names from Geonames as synonyms.

Klein et al.[14] describe a framework for semi-automatic creation of a historical commodities thesaurus from a controlled vocabulary. Experts manually compile concepts with preferred and alternative labels from a collection of archived ledger books and map them to corresponding DBpedia resources. These resources are used to construct a simple SKOS thesaurus. Next, the system follows the links to DBpedia to obtain category information, which is introduced as new broader concepts. They further enrich the thesaurus by acquiring sibling concepts from DBpedia categories.

2.3 Word Sense Disambiguation

Word sense disambiguation is the task of determining the sense of a word in a specific context. It can be viewed as a classification task, in which some automatic method is used to assign words to a sense (class) given the context and/or information from an external knowledge source. These knowledge sources can be structured (dictionaries, thesauri, ontologies) or unstructured, like raw or sense-annotated corpora and collocation resources. A particular kind of approaches use graph structures for disambiguation. Usually, those graphs follow the idea of a *lexical chain*[15, 16, 17]. Lexical chains are sequences of words, which are semantically related to their successors.

Navigli and Velardi[18] have proposed the Structural Semantic Interconnections (SSI)

algorithm. It features the development of a context-free grammar for lexical chains using often encountered relation patterns in WordNet (semantic interconnections). For each monosemous word in a context, SSI builds semantic graphs from WordNet. In an iterative fashion, ambiguous words are then mapped to senses in the semantic graphs. The measure for this is based on a weighting of semantic interconnections from the grammar.

Another work of Navigli and Lapata[19] investigates the usage of common graph connectivity measures on WordNet subgraphs. For a sentence, whose content words (nouns, verbs, adjectives, adverbs) have to be assigned a sense, their algorithm selects the most appropriate senses by finding the corresponding lexical chain from WordNet. Beginning with each candidate sense of the first word in the sentence, they perform a depth-first search to construct the graph, which contains all possible paths between all possible senses for each content word. Afterwards, each vertex is ranked according to some connectivity measure and the highest ranking candidate sense is selected for each word. Navigli and Lapata found degree centrality and PageRank[20] to be the best performing connectivity measures allowing their algorithm to reach state-of-the-art.

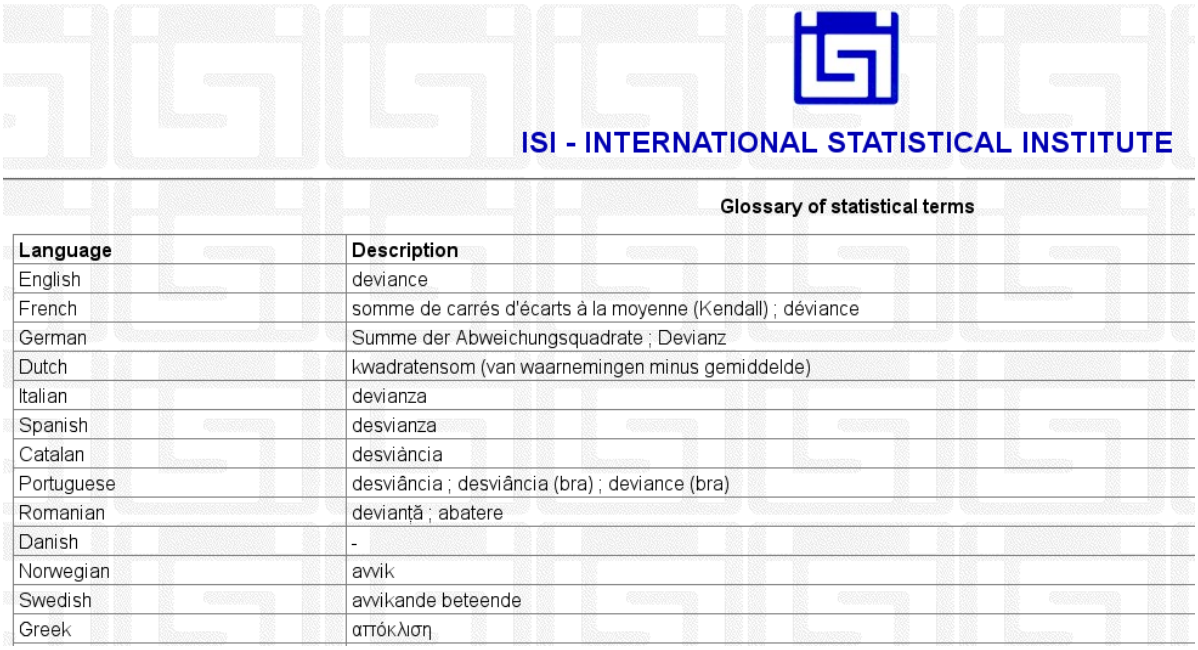
3 Extraction of the Model

3.1 Mining the Online Glossary

The ISI Multilingual Glossary of Statistical Terms Due to the focus of the Interlingua project, the ISI Glossary of Statistical Terms[4] was chosen as the source for model extraction. It is provided by the International Statistical Institute whose mission is "to promote the understanding, development and good practice of statistics worldwide"[5]. One action in its agenda was the compilation of a comprehensive multilingual dictionary of statistical terms. Starting 1993, an international team of specialists gathered 3564 items, many of them with several synonyms, in 31 languages.

Mining The online glossary contains one web page per distinct word sense showing a table with one column *Language* and one column *Description*. The entries for the language column contain the 31 languages, due to the focus of the Interlingua project, only English, German and French are considered in this thesis. The entries for the description column are the synsets for the given language, i.e. sets whose members are valid, semantically equivalent translations of the sense (in the following simply referred to as synonyms). Since the charts are implemented as basic HTML tables, one can obtain the synsets for each language just by accessing the Document Object Models of the corresponding pages. Furthermore, the URLs of the glossary entries follow an incremental pattern (*http://isi.cbs.nl/glossary/termX.htm* where X is an integer greater than zero), which allows for a very straightforward approach to mining the content of the glossary: while incrementing X until getting a 404 HTTP response code, parse the

[4] http://isi.cbs.nl/glossary/index.htm
[5] http://www.isi-web.org/about-isi/objectives-mission



ISI - INTERNATIONAL STATISTICAL INSTITUTE

Glossary of statistical terms

Language	Description
English	deviance
French	somme de carrés d'écarts à la moyenne (Kendall) ; déviance
German	Summe der Abweichungsquadrate ; Devianz
Dutch	kwadratensom (van waarnemingen minus gemiddelde)
Italian	devianza
Spanish	desvianza
Catalan	desviància
Portuguese	desviância ; desviância (bra) ; deviance (bra)
Romanian	devianță ; abatere
Danish	-
Norwegian	avvik
Swedish	avvikande beteende
Greek	αποκλιση

Figure 2: Screenshot of the ISI glossary page http://isi.cbs.nl/glossary/term933.htm

HTML of the page, and pick the English, German and French synsets from the table. After processing the glossary, one obtains a list holding a mapping from languages to synsets for each word sense.

Auxiliary Issues Sometimes, the synonyms are syntactically ambiguous, i.e. they contain bracket expressions, examples being *(bedingte) Erfassung* and *cluster (point) process*. Here, the braces indicate an optional addition to the main term and the ISI glossary treats the whole expression as one synonym, though *bedingte Erfassung* and *Erfassung* resp. *cluster point process* and *cluster process* can be viewed as two different ones. Conveniently, this perspective also eliminates the braces, which is a preprocessing step for the Entity Linking task. Therefore, a small subroutine has been added to the mining algorithm for resolving bracket expressions in the synsets.

According to ISI, some of those specialists were tasked with checking the quality of the translations, they also acknowledge the possibility of wrong ones and are open to comments and corrections. During the work on this thesis, two types of dubious translations have been noticed, the first one being word for word ones not fitting the statistical context (e.g. *Baumbeschneidung* as the German equivalent for *tree-pruning*) and the second one being attempts to render neologisms into another language, e.g. *Klumpen letzter Ordnung* as the German equivalent of *ultimate cluster*. Though technically correct, its value is debatable.[6]

[6] A Google search for *"ultimate cluster"* or *"ultimate cluster" statistics* gives some results in English which define the term, but no translations. The twelve search results for *"Klumpen letzter Ordnung"* are either irrelevant or seem to be accurate copies of the corresponding ISI Glossary entry.

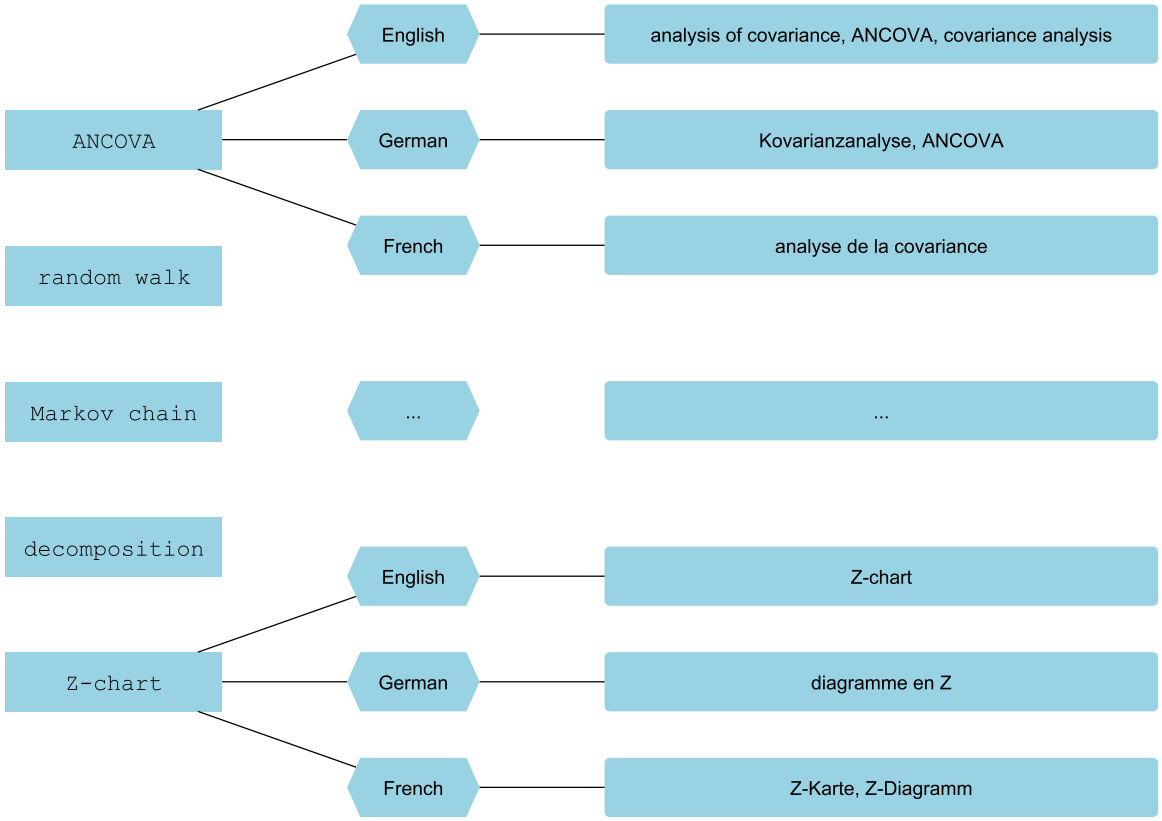


Figure 3: The intermediate representation of the ISI glossary. Arbitrary synonyms have been chosen as mapping identifiers for the purpose of illustration.

The problems with both types of dubious translations become evident at a later stage during the semantic enrichment. When querying web resources for information in the case of word for word translations, erroneous data may be introduced to the model, since the word senses do not match. In the case of a neologism, a worst, an average and a best case can be distinguished. The worst case is the same as it would be for a word for word translation. Observation has shown that many neologisms (as well as some very specialized terms as well) have been linked to broader concepts, this can be considered as an average case. Arguably, the best case is the absence of any result, as this is the only alternative which does not lead to noise in the model.

3.2 Building the SKOS Model

For representing the multilingual domain model in a machine-readable way, the *Simple Knowledge Organization System (SKOS)* has been chosen, which is the W3C recommendation for modeling and linking knowledge organization systems for the Web[21].

In library and information sciences, knowledge organization systems (KOS) denote tools for organizing large collections of objects, which may be books, museum artifacts

but also pieces of information on some specific topic or domain. Example kinds of KOS are taxonomies, in which objects are modeled using hierarchical relations, and thesauri, which additionally feature associative relations as well as synonymy. Since these share many of their properties, it was possible to develop SKOS as the first data model able to capture them in a standardized fashion.

Following Tim Berners-Lee's vision of "a web of data that can be processed directly and indirectly by machines"[22], the W3C's Semantic Web Activity[23] has been directed at developing the foundational standards for the Semantic Web, being the Resource Description Framework (RDF) as a data abstraction and syntax, the RDF Vocabulary Description Language (RDFS) and the Web Ontology Language (OWL) together as a data modeling language and the SPARQL Query Language and Protocol as means for interacting with data. Applying those technologies across applications enables software agents to reason about information from different sources and provide for intelligent services.

However, they depend on extensive knowledge bases, of which larger parts cannot be compiled automatically. Ontology engineering is a time consuming process, requiring in-depth understanding of a domain as well as expertise in formal modeling and logical forms of expression. Therefore it is desirable to use more informal but still profound experience and best practices in the library and information sciences, as many KOS are already made available on the Web[21]. The aim of SKOS now is to bring together these communities and the Semantic Web by facilitating the migration of existing KOS to the Semantic Web as well as the creation of machine-readable knowledge bases in cases, in which there is no need for the formal semantics inherited by heavyweight ontologies. Such a migration from a plain representation to SKOS has been e.g. successfully accomplished for the United Nations Food and Agriculture Organization's AGROVOC Thesaurus in 2009 by Caracciolo et al. after already having tried remodeling in OWL. Their experience with porting a thesaurus to OWL was that the primarily terminological nature of thesauri is hardly compatible with the logical rigor required by OWL[24]. Formal ontologies express facts about the world by asserting axioms and modeling the formal relationships between them. Thesauri on the other hand describe sets of meanings through natural language. While they can be arranged in hierarchies and association networks, these do not have any formal semantics since they are intended as intuitive means for mapping subject domains.

Though the goal of the framework presented in this thesis is the creation of an ontology, SKOS is still the more suitable choice for the representation of the domain model than OWL. Since the base knowledge source of the extracted model is a glossary, SKOS is the more fitting choice for its terminological nature. Also in the enrichment stage, lexical information is of great importance. Furthermore, the framework in its current state does not aim for the capture of formal constraints, which would make the usage of OWL indispensable.

From the aforementioned intermediate representation, a SKOS representation of the glossary is built and serialized in RDF/XML. The SKOS API[7] is employed for this purpose, which provides methods for simple creation and access of elements in the SKOS namespace. By means of using *belief network* as an example, the following section explains this procedure and the employed SKOS namespace elements as well as SKOS properties used during enrichment at a later stage.

<Belief_network> rdf:type ***skos:Concept*** .

The basic unit of the SKOS data model is a *concept* (denoted as *skos:Concept*, *skos* is the namespace prefix for the domain http://www.w3.org/2004/02/skos/core#). A concept is identified by a URI, making it uniquely identifiable and referable in the World Wide Web. It is intended to represent an abstract idea or notion, a unit of thought, and should be applied to model one specific meaning in a KOS. Accordingly, a concept is created for each element of the intermediate representation list.

<Statistics> rdf:type ***skos:ConceptScheme*** .
<Belief_network> ***skos:inScheme*** *<Statistics>* .

Concepts can be aggregated in a *concept scheme* (*skos:ConceptScheme*), which is also identified by a URI. Though it is intended to represent one KOS, concepts are not restricted to the membership in exactly one concept scheme. There are also no mechanisms to state that only some certain concepts belong to a scheme. During conversion of the mined information from the glossary to SKOS, one concept scheme is created to capture the domain of statistics.

The property *skos:inScheme* is used to assign a concept to a concept scheme. However, only the range of skos:inScheme is explicitly stated, which means that any resource can be member of a scheme.

<Belief_network>
skos:prefLabel *"belief network"@en ;*
skos:altLabel *"Bayesian network"@en ;*
skos:prefLabel *"Bayessches Netzwerk"@de ;*
skos:prefLabel *"réseau bayésien"@fr .*

Concepts can have lexical labels describing their meaning in natural language. There are several kinds of labels, two of them being preferred (*skos:prefLabel*) and alternative ones (*skos:altLabel*). Lexical labels consist of a string of UNICODE characters and an optional language tag as defined in [25]. No domain is stated for skos:prefLabel and skos:altLabel, therefore any resource can be labeled with SKOS lexical labels. Since skos:prefLabel and skos:altLabel are disjoint properties, the same label cannot be both preferred and alternative. Furthermore, a concept can only have one preferred label per language at most.

SKOS lexical labels are the most important feature for the presented framework to capture the terminological nature of the domain model. The annotation with language tags allows for the explicit modeling of multilinguality in concepts. The option to assert

[7] http://github.com/simonjupp/java-skos-api

several labels is vital for the notion of synonymy. In the example, one can see that the concept <Belief_network> has translations in English, German and French and that for English, there are two synonyms which can be used interchangeably. I.e. the language-synset map can be represented one-to-one.

<Belief_network> ***skos:definition***
"A Bayesian network or belief network is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph."@en .

Documentation properties are used to provide information about concepts. Neither domain nor range are stated, so there are no restrictions on the information assigned to a resource: it could be a definition in plain text, editorial information in some markup language or an image. The information can also be annotated with a language tag. This enables us to assert different definitions in individual languages for a concept during the enrichment stage.

<Belief_network> ***skos:broader*** *<Networks> .*

Concepts can be linked with two categories of semantic relations, hierarchical and associative. *skos:broader* and *skos:narrower* are used to assert (by convention) direct hierarchical links between two concepts, *skos:related* introduces an associative relation; skos:semanticRelation is their top property. skos:broader and skos:narrower are inverse to each other, skos:related is symmetric. The hierarchical properties are disjoint with skos:related. It is important to note that nothing is stated on reflexivity and transitivity of these relations. Semantic relations are obtained during the enrichment of the model with information from DBpedia.

These few steps are enough to provide a simple machine-readable representation in SKOS of a given input glossary. At this stage, the model can already be used as the knowledge base of a dictionary.

	Total no. of concepts	No. of concepts with synonyms
English	3561	501
German	3492	627
French	3464	695
All three	3433	200
Only English	46	

Table 1: Number of extracted concepts per language

The left colum of Table 1 indicates the number of extracted concepts, which have at least one label for the given language. Its values are compared to the number of extracted concepts, which have more than one (i.e. one preferred and at least one alternative) label per language. Since, as mentioned before, English preferred labels

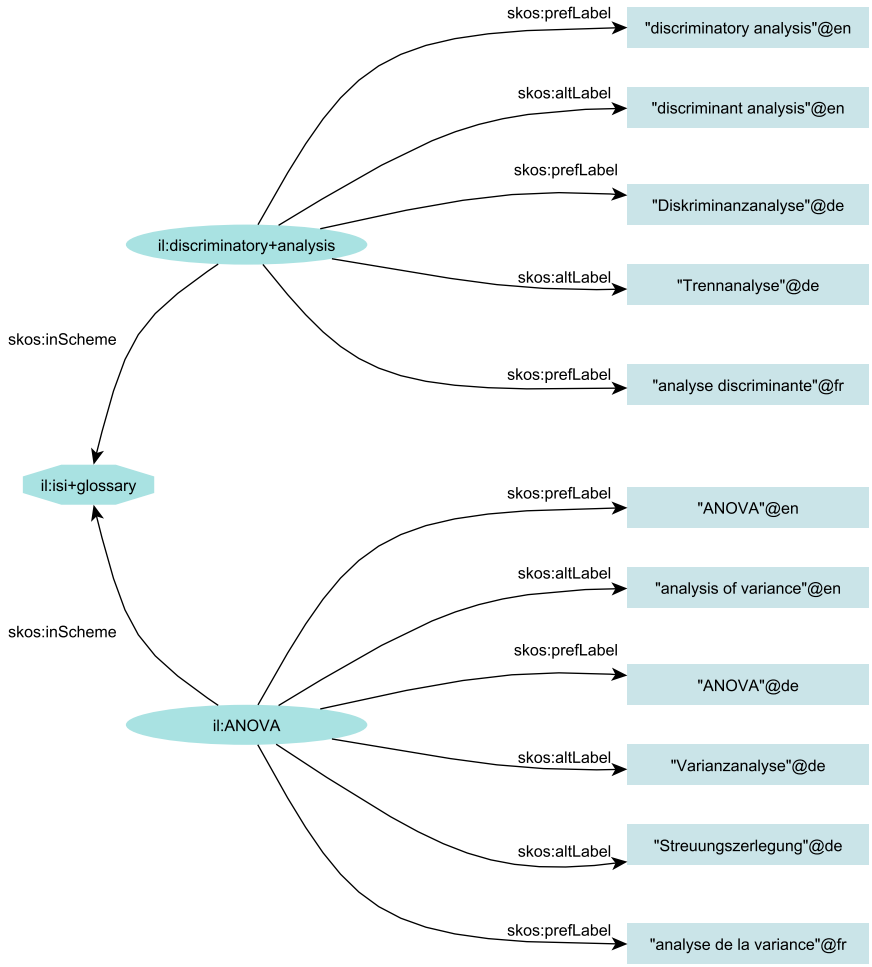


Figure 4: The SKOS representation of the extracted model

can be used for identification, the number of concepts with English labels corresponds to the total number of concepts. The mismatch between the number of entries in the ISI glossary (3564) and the number of extracted concepts (3561) comes from duplicate information being collapsed into one concept. Missing translations can be obtained during enrichment, but due to a rather high multilingual coverage of the ISI glossary, such a procedure has not been implemented. It can be considered for future work. On the other hand, 3433 concepts have labels in all three languages, 200 out them with synonyms for each one. These can be considered the best class of extracted concepts.

4 Semantic Model Enrichment

After extracting data from a source like the ISI glossary, one has a fitting base for learning a comprehensive, multilingual domain model in a way similar to [14]. In the presented framework, enrichment means the discovery of semantic relations in the sense of Ontology Learning on the one hand, and textual information like concept definitions on the other hand, as well as their introduction to the model. DBpedia[8] is used as a prominent linked open dataset for retrieval of information on concepts in the model.

4.1 DBpedia

DBpedia is a crowd-sourced community effort to extract knowledge from Wikipedia and make it freely available using Semantic Web standards [8]. Every resource on DBpedia is automatically mined from a Wikipedia article, the information is then structured according to a community maintained ontology. Started in 2006, the project has since grown to be a central hub of Linked Open Data with over 27 million outgoing links to over 30 datasets and over 39 million incoming links from over 240 datasets [8].

With respect to evaluation of the presented enrichment approach, it is interesting to consider the multilingual nature of DBpedia. There are localized versions in 125 languages corresponding to the localized versions of Wikipedia. All these versions together describe 38.3 million resources. For each language, there is also a canonicalized data set only containing resources, which have an equivalent in the English DBpedia.[8][9]

	Localized Data	Canonicalized Data
English	4 584 616	4 584 616
German	1 692 634	857 196
French	1 504 453	942 505

Table 2: Number of resources in different languages

Unfortunately, DBpedia does not provide numbers for the overlap between individual languages. It is still interesting to observe the disparity between German and French. It hints at a higher effort for multilingual interlinking in the French community than in the German one.

[8] http://wiki.dbpedia.org/services-resources/datasets/dbpedia-data-set-2014
[9] http://dbpedia.org/services-resources/datasets/dataset-statistics

4.1.1 DBpedia Metadata

Resources on DBpedia are annotated with metadata through a set of distinct properties in the DBpedia namespace as well as the Dublin Core initiative[10] and the FOAF vocabulary[11] (among others). This section highlights those properties, which are relevant to the presented framework.

dbpedia:Bayesian_network ***rdfs:label*** *"Bayesian network"@en .*

rdfs:label is used to denote the title of the corresponding Wikipedia article. The asserted values can also be seen as lexical representations of resource URIs. So, the primary way to link model concepts to DBpedia resources is to query DBpedia's SPARQL endpoint for resources with same labels as the concepts.

dbpedia:Bayesian_network ***dbpedia-owl:abstract*** *"A Bayesian network or belief network is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph."@en .*

Abstracts of Wikipedia articles are asserted to a resource via dbpedia-owl:abstract. They are used in the enriched model as concept definitions.

dbpedia:Bayesian_network ***dbpedia-owl:wikiPageWikiLink*** *dbpedia:Statistical_model .*

In Wikipedia, articles are unidirectionally connected by links embedded in the text. DBpedia mirrors this structure with the dbpedia-owl:wikiPageWikiLink property. One can observe, that in Wikipedia, articles are linked as term explanations, related reading or for comparison of different ideas. This leads to the assumption of underlying semantic relations in Wikipedia links. Therefore, they are used to introduce relations between concepts in the enriched model.

dbpedia:Bayesian_network ***dcterms:subject*** *dbpedia:Category:Networks .*

dcterms:subject is used in DBpedia to assign resources to categories. They correspond to categories in Wikipedia and are analogously arranged in hierarchies with the skos:broader property. During enrichment, taxonomic relations are introduced to the model based on DBpedia categories.

dbpedia:Bayesian_network ***foaf:isPrimaryTopicOf*** *wikipedia:Bayesian_network .*

foaf:isPrimaryTopicOf indicates the Wikipedia article to which the resource corresponds. These assertions are reused to denote the information source of the enrichment. The FOAF vocabulary specification defines this property as the relation between a resource and a document whose primary topic is this resource.

[10] http://dublincore.org/documents/dcmi-terms/
[11] http://xmlns.com/foaf/spec/

dbepdia:Elimination
dbpedia-owl:wikiPageDisambiguates *dbpedia:Gaussian_elimination ;*
dbpedia-owl:wikiPageDisambiguates *dbpedia:Elimination_tournament .*

For ambiguous terms like "Elimination" and "Regression", Wikipedia does not contain a single article but instead a *disambiguation page*. There, several articles corresponding to distinct word senses of the term are suggested. DBpedia models this behavior with dbpedia-owl:wikiPageDisambiguates assertions, which link to resources describing distinct word senses. Ambiguous DBpedia resources are central to the presented framework, since they induce the disambiguation mechanism.

4.1.2 Why DBpedia?

Besides DBpedia, there are several other Linked Open Data sets, which have been successfully used in research, like Freebase[26] and OpenCyc[11]. Out of them, DBpedia is closest to Wikipedia and thus benefits from the largest open community curating a knowledge source, widely established outside research too. But there are also concrete downsides to other data sets with respect to this thesis' focus.

Similarly to DBpedia, the ground of Freebase is derived from Wikipedia and allows content editing by users in addition. Like the former, Freebase arranges more popular information in a community maintained ontology, which allows for fine-grained semantic queries. Though for more specialized concepts in statistics, such granularity is absent like in DBpedia, DBpedia's main goal leads to an advantage at this point. Due to its rigorous mirroring of Wikipedia, links between resources are present even for highly specialized resources. Freebase lacks such links, e.g. there are no links for the central limit theorem[12]. This is a grave obstacle for the presented approach concerning enrichment of semantic relations. Furthermore, there are technical considerations in respect of Freebase. Instead of SPARQL, which is the standard and W3C recommendation for interaction with semantic data, it uses the Metaweb Query Language (MQL) which has been developed specifically for Freebase[26]. As a consequence, integration of Freebase into common APIs like Apache Jena[13] is impeded. Moreover, the Freebase API has been deprecated and activity on Freebase retired[14].

OpenCyc[12] is a manually assembled ontology, aimed at capturing common-sense knowledge. The nature of OpenCyc would be the main obstacle for the presented framework. Resources are formulated as axioms and assertions in the language CycL, which is based on first-order logic. As mentioned earlier, such a formal approach is hardly compatible with the terminological nature of the model in the presented framework.

[12] http://www.freebase.com/m/09t70
[13] http://jena.apache.org/
[14] http://plus.google.com/109936836907132434202/posts/bu3z2wVqcQc

4.2 Enrichment

4.2.1 Querying DBpedia

For enrichment of the extracted model, the presented framework links its concepts to resources in DBpedia. The basic approach for this consists of querying DBpedia's SPARQL endpoint[15] for resources with same labels as the annotated concept labels.

Since the English DBpedia contains the most data, its SPARQL endpoint is used as the primary entry. For each concept in the model, the English SPARQL endpoint is queried for resources whose rdfs:label matches one of the annotated skos:prefLabel or skos:altLabel. Because equivalent resources across localized versions of DBpedia are asserted as owl:sameAs, by establishing a link to an English resource, such assertions can be resolved to obtain the corresponding German and French resources (issues arising during resolution will be highlighted later on). If no English resource can be found, the German and French SPARQL endpoints are queried manually to cover cases of non-canonicalized data.

In the early stage of work, an assumption was used based on the discrimination between preferred and alternative labels. Since preferred labels can be seen as the main lexicalization of a concept, it would suffice to query DBpedia only for preferred labels first. If no resource can be obtained this way, retry with alternative labels. This approach turned out to result in a certain loss of information, because in some cases, one can link a concept to multiple resources based on different labels. Observations have shown, that the results for alternative labels are sometimes more fitting with respect to the domain of statistics. Thus the decision was taken to allow one-to-many linking in this stage of enrichment, resolving such ambiguity will be approached in future work.

After having found corresponding multilingual DBpedia resources for the concepts, one can retrieve the enrichment metadata by querying the resources for the properties mentioned above. It consists of respective Wikipedia URLs, abstracts in English, German and French, links to related resources and the categories to which the resource belongs. For categories, a breadth-first search like procedure is performed. Once categories for all concepts are mined, DBpedia is queried for their parents. In case of cyclic hierarchies, which mostly appear on more abstract levels, expansion for a resource is stopped before entering the cycle (dbpedia:Thought → dbpedia:Mind → dbpedia:Concepts_in_metaphysics → dbpedia:Philosophical_concepts → dbpedia:Philosophy → ~~dbpedia:Thought~~). This is repeated until no more new categories can be obtained.

On reaching an ambiguous resource, i.e. a resource with dbpedia-owl:wikiPageDisambiguates assertions, the suggested resource URIs are stored as identifiers of candidates for the correct concept sense. When enrichment has been applied to all monosemous concepts, ambiguous resources are processed by the disambiguation procedure. The resources determined to be representing correct concept senses are then mined for metadata.

[15] http://dbpedia.org/sparql

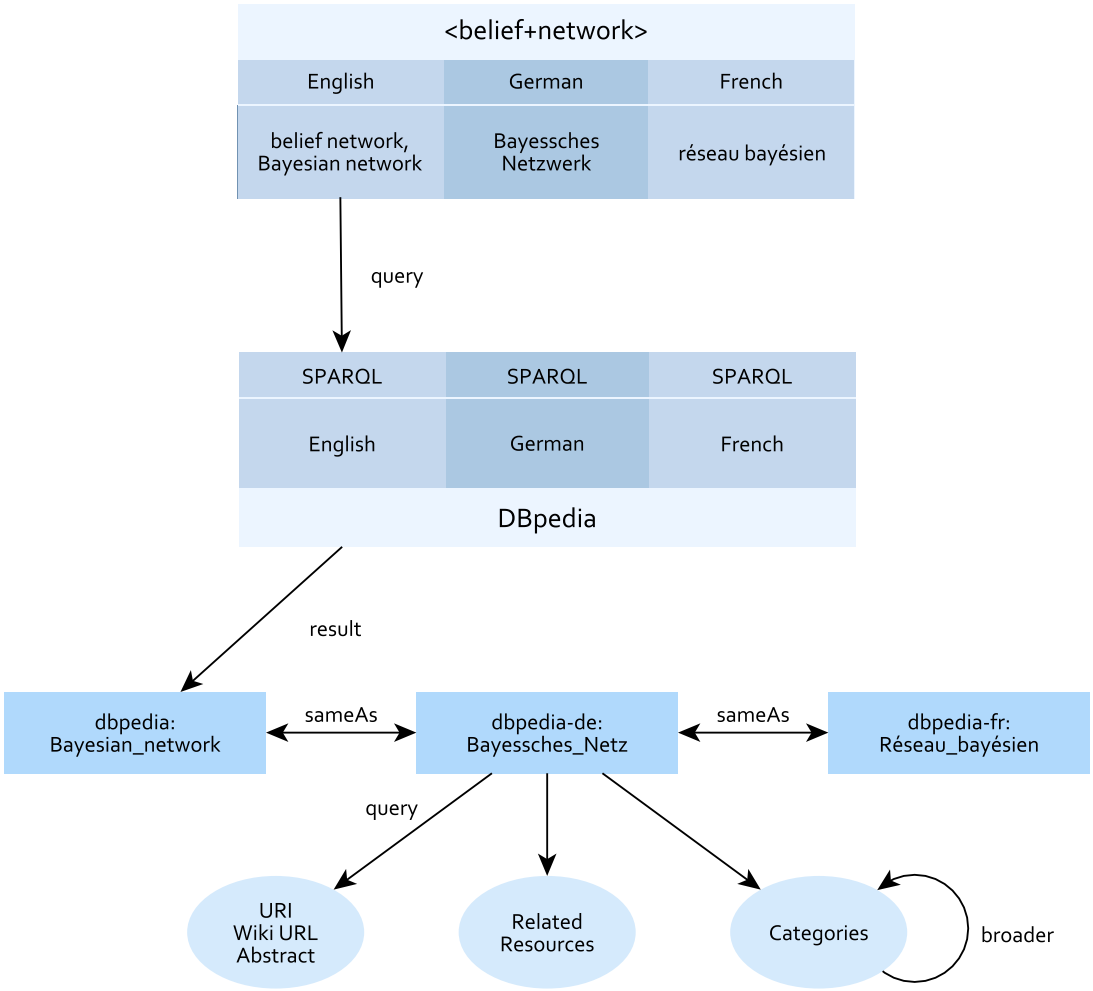


Figure 5: This example illustrates how information on the concept *belief network* is retrieved. Querying the English SPARQL endpoint for resources whose labels match the English concept label results in *dbpedia:Bayesian_network*. The corresponding German and French resources are obtained through resolving the sameAs links.

4.2.2 Adding Retrieved Information to the Model

Having mined metadata from DBpedia for all concepts, enrichment is applied to the model. To store the linking between a concept and a resource, the DBpedia URIs for each language are annotated to the concept with corresponding language tag. Dublin Core's *source* (dcterms:source) property[16] is used here, indicating that the concept has been partially derived from DBpedia. The same is done for the Wikipedia URLs with the help of foaf:isPrimaryTopicOf and for abstracts with skos:definition. By enriching

[16]http://purl.org/dc/terms/source

concepts with definitions, the framework learns new lexical information about the earlier extracted model.

For links to related resources, the related DBpedia URIs have to be discovered in the model. After all concept-DBpedia links have been established in the model, it is queried for concepts, whose source annotations match the related URIs. If a match is found, a skos:related assertion between the corresponding concepts is introduced at the concept, which has the outgoing link. One should note, that though by convention, skos:related is a symmetric property, the assertion is derived from unidirectional links. From the perspective of a semantic relation, this is reasonable, since unidirectional links can be interpreted as associative *mentions* and *is-mentioned-by* relations. In the context of Ontology Learning, this step constitutes the learning of non-taxonomic relations. DBpedia URIs for which no match can be found are not discarded, but stored separately. They are central to the graph-based disambiguation approach.

When applying category enrichment, the framework first checks for correspondence between mined categories and concepts, assuming that if a concept can not be linked to a proper DBpedia resource, it can still be equivalent to a category. The model is queried for concepts whose labels match the categories. If a matching concept is found which is not enriched, the URI of the category is annotated via dcterms:source, thus linking the concept. If the match already has an established linking, the framework stores it internally as the corresponding concept for the category. Else a new concept is introduced with the category URI as its source. After having linked the categories to concepts, according hierarchic relations are asserted with skos:broader. By this means, a taxonomy is learned for the concepts and introduced to the initially flat model.

4.3 Arising Issues

Several problems arise during enrichment when querying DBpedia.

As mentioned above, there are difficulties in the resolution of multilingual owl:sameAs links. Cabrio et al. indicate in [27] that in some cases, owl:sameAs links are missing between corresponding resources on different versions of DBpedia. This can happen due to misalignment in localized versions of Wikipedia. E.g. *Mean* and *Average* are two separate articles on the English Wikipedia, but are subsumed in the German article *Mittelwert*, which is only linked to *Average*. Consequently, DBpedia shows the same behavior and dbpedia:Mean has no owl:sameAs link to dbpedia-de:Mittelwert. Because of such cases, one can not assume DBpedia's owl:sameAs resolution to be complete. As a fallback in this example, the framework queries the German endpoint after having obtained both English and French resources from the English endpoint.

Furthermore, the English DBpedia appears to completely lack dbpedia-owl:wikiPageWikiLink assertions, i.e. links to related resources. While some related resources can be gained from the German and French DBpedia, it is unclear how many relations the framework misses. Since the English DBpedia and Wikipedia contain a substantially larger number of resources than their German and French counterparts it is reasonable to assume, that they also contain substantially more relation links. This causes a specific worst case for the enrichment: if a concept can only be linked to an English DBpedia resource, no rela-

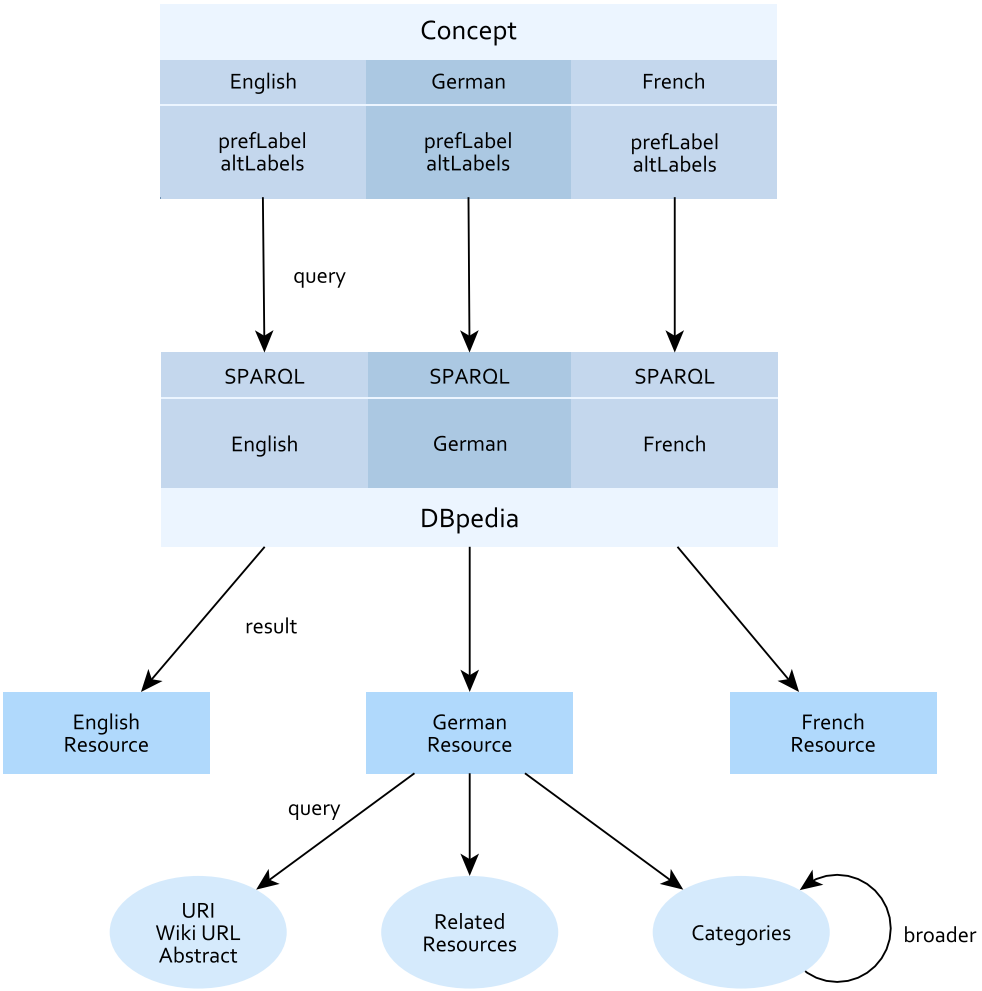


Figure 6: General procedure for circumventing missing sameAs links. The multilingual resources are obtained through three separate queries in contrast to a single one (see Figure 5 on page 23).

tions can be learned. Extracting such missing links using Wikipedia will be approached in future work.

These two issues could be observed during the work on this thesis. However, research on inconsistency detection in DBpedia has recognized several more forms of errors. Töpper et al.[28] identify three classes, syntactic, logical and semantic errors. Syntactic errors are mainly caused by syntax errors in RDF generation but do not propagate errors into reasoning due to RDF parsers. Logical errors are contradicting assertions, usually introduced through misalignment of the DBpedia ontology and Wikipedia content. Semantic errors are erroneous content. Wienand and Paulheim[29] give an extensive treatment on detection of incorrect numerical data in DBpedia as well as their causes.

4.4 Evaluation

	No. of linked concepts
English	763
German	535
French	460
All three	334
Distinct concepts	1028

Table 3: Number of concepts linked monosemously to DBpedia per language

According to expectations, the relation between the numbers of achieved concept-DBpedia links for the three languages corresponds to the relations between language labels in the extracted model and between resources across localized DBpedia. These numbers however only represent monosemous cases, i.e. links to DBpedia resources, which do not have to be disambiguated. Evaluation of the disambiguation procedure is shown in the next chapter.

	No. of relations	No. of related concepts
Associative	2016	552
Hierarchic	228	165
Introduced categories		204

Table 4: Number of monosemously learned relations and categories

In contrast to 763 concepts with English enrichment, the number of 552 concepts, for which associative relations can be learned, is caused by the aforementioned lack of information on Wikipedia links on the English DBpedia. So, the average number of 3.65 associative relations for each concept does not reflect the full knowledge of Wikipedia.

5 Named Entity Disambiguation

As mentioned in the chapter on Semantic Model Enrichment, the problem of ambiguity arises during the linking of model concepts to resources from DBpedia. Disambiguation pages mark terms with multiple senses and suggest resources which disambiguate them. Therefore, to achieve an as exhaustive as possible linkage of concepts and resources, the task of Named Entity Disambiguation has to be solved, i.e. the word sense disambiguation of named entities in a knowledge base during linking.

For this purpose, a multi-layered algorithm leveraging domain knowledge is presented which combines different techniques to maximize performance. Its subroutines consist of a graph-based, a corpus-based and a model-based approach. Due to their modular

nature, the algorithm can be arranged in a two-pass or three-pass way. In a two-pass setting, the graph-based procedure is followed by either the corpus-based or the model-based disambiguation. The three-pass alternative applies the graph, the corpus and the model in this particular order. Each step disambiguates a certain number of DBpedia resources, the input for the next pass are the resources, which could not have been resolved in the current one.

The corpus and model approaches are based on the vector space model, which represents text documents as vectors. For the aforementioned methods, the vectors hold term frequencies, i.e. the frequencies of each term appearing in the considered documents. Using cosine similarity, these representations can be compared. A smaller angle between two vectors (i.e. higher cosine value) indicates higher similarity between the corresponding documents. Bunescu and Pasca have shown the applicability of cosine similarity for word sense disambiguation in [30].

5.1 A Graph-based Approach

As discussed in chapter 4, DBpedia, as a linked open dataset, not only provides information on specific resources, but also on the relations between different entities. When considering a certain domain like probability theory and mathematical statistics, those can be used to learn its structure. This section describes a method to infer domain membership of DBpedia resources, i.e. disambiguate them, based on structural knowledge of the domain.

The enrichment stage of the presented framework learns semantic relations between the concepts in the model. Using the information on associativity and hierarchy, a graph can be constructed with enriched concepts as nodes and links as undirected edges.[17] Additionally, the graph incorporates the mined links to DBpedia resources, which do not match any concept in the model, as well as all candidate resources for disambiguation.

One can now assume a certain kind of clustering. Correct candidates, i.e. DBpedia resources describing an idea in probability theory or mathematical statistics, are related to other resources in the target domain. Thus they belong to the main cluster, a dense sub-graph. In contrast, irrelevant resources are isolated since they do not belong to the domain. The isolated candidates can be pruned, the interlinked ones are then assigned as correct word senses for the corresponding concepts. Here, it is important to have the additional links from DBpedia. In many cases of ambiguity, it can be observed that the set of non-matched resources contains the correct word sense candidates.

This results in following classification rule for ambiguous resources: For one ambiguous resources, consider each node representing a candidate resource. If no node representing an unambiguous resource can be reached from the candidate node, classify it as irrelevant. Otherwise, mark it as the correct word sense.

As mentioned earlier, many potential relations cannot be learned due to the lack of resource linkage on the English DBpedia. The constructed graph therefore can not be

[17] Similarly to the symmetric treatment of Wikipedia links in terms of semantics, hierarchic relations can also be considered bidirectional.

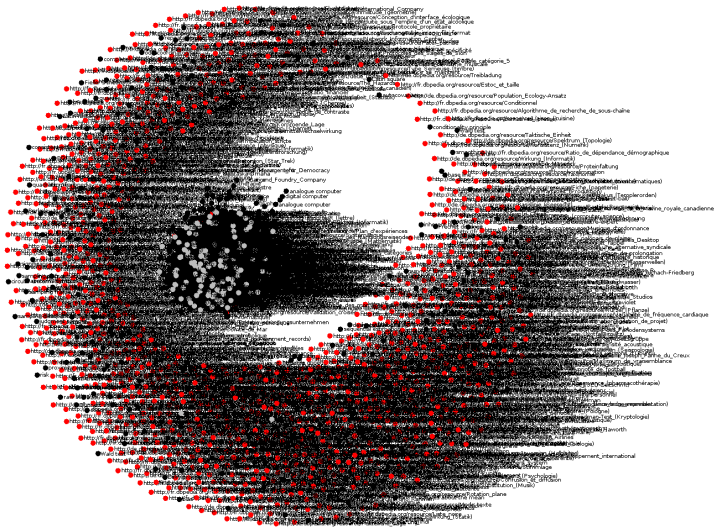


Figure 7: Full graph with interlinked cluster and isolated nodes. Black nodes correspond to enriched concepts, gray ones are additional linked resources from DBpedia. Red nodes are candidates resources for disambiguation.

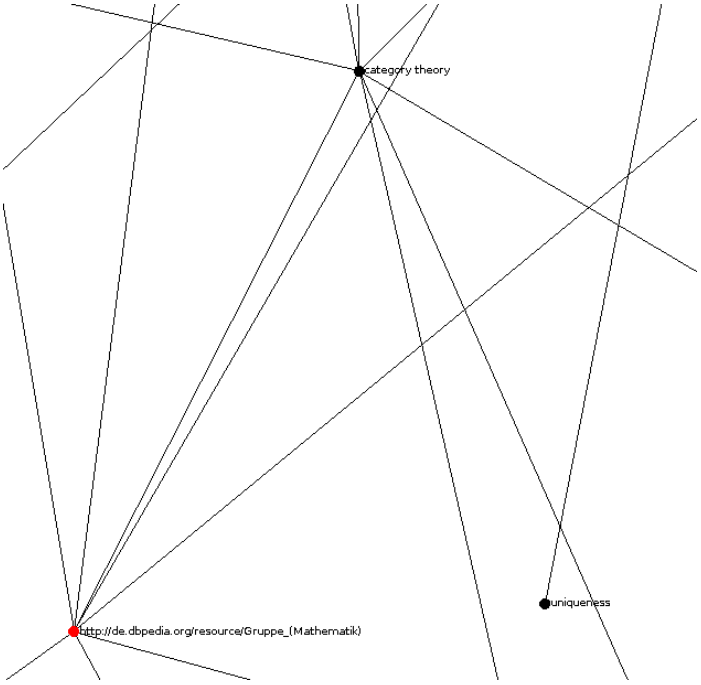


Figure 8: *http://de.dbpedia.org/resource/Gruppe_(Mathematik)* is a candidate for the ambiguous concept *group*. Due to its semantic link with *category theory*, it belongs to the target domain, in contrast to other candidates describing social group or newsgroup.

assumed to be exhaustive in terms of domain structure. For the presented approach, this means that not all correct word senses can be found in the main cluster, some are isolated even though they belong to the domain. Disambiguation therefore depends on at least one further procedure to capture those cases. Nevertheless, the graph-based approach has been able to disambiguate 136 out of 275 ambiguous resources in total without error. In each case covered by the graph, only one of the candidates appears in the domain cluster.

5.2 Corpus-based Disambiguation

Due to the focus of the Interlingua project, there is access to a multilingual corpus of documents consisting of textbooks in English, German and French. Since these provide definitions for many terms described by the ontology, it is possible to leverage this additional textual domain knowledge.

During the work on the project, a mapping has been developed between the ontology and the corpus as illustrated in Figure 9. Terms from the book indices are matched against concept labels in the corresponding language. Once such a linking is established, one can follow the page numbers in the indices to find introductory sections for the terms. More advanced information extraction techniques can be used to narrow the location of the actual introductions down to particular segments instead of whole pages.

The hereby obtained sections can be utilized for disambiguation based on the assumption, that texts covering the same topic apply similar language. Ideas from varying domains on the other hand would differ in that regard. For each concept which is linked in the aforementioned manner to an index term in a certain language, one can derive a term vector representation of the corresponding segment in the corpus. The candidate resources from DBpedia also provide textual information, namely short abstracts, as well as full Wikipedia articles. Either one of those two types of text can be transformed to a vector representation and used for the computation of cosine similarity with respect to the vector obtained from the corpus. The candidate which achieves the highest cosine similarity, i.e. has the greatest collocational overlap, is then determined to be the resource representing the correct word sense.

5.3 Model-based Disambiguation

Since the extracted and enriched model is motivated as being a source of domain knowledge, one can conclude its value for domain-dependent reasoning. Concerning ambiguity in DBpedia, it is interesting to evaluate the feasibility of inferring domain membership of DBpedia resources, i.e. disambiguation, with the help of the lexical information contained in the model. Similar to the concept-based approach, one assumes that documents describing ideas from a certain field use some common, standardized terminology.

As illustrated in Figure 10, the concept labels are used to build queries represented as vectors. There are three queries, one for each English, German and French, encoding the entire terminological knowledge of the model. Before the actual disambiguation procedure, the labels in all three languages of every concept are added as a disjunction

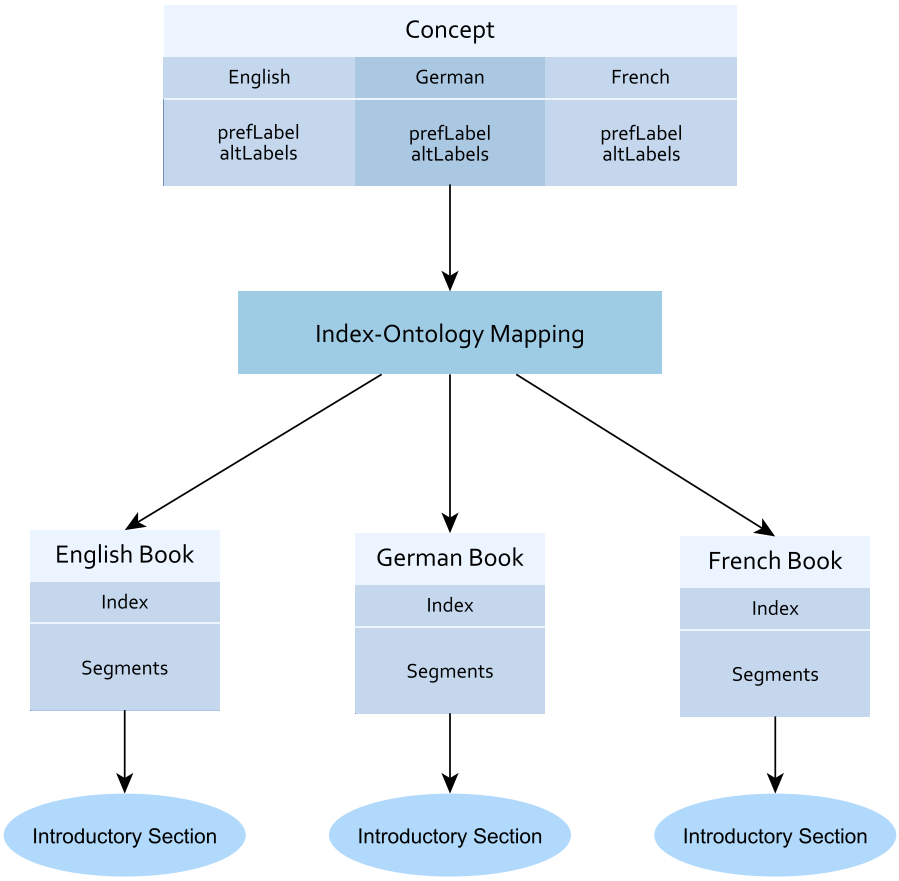


Figure 9: Retrieval of introductory sections for concepts.

to the corresponding query. To resolve ambiguity of DBpedia resources, the term vectors of either the abstracts or full article texts of each candidate in a certain language are compared to the corresponding query. The candidate which achieves the highest cosine similarity, i.e. features the most queried terms, is then determined to be the resource representing the correct word sense.

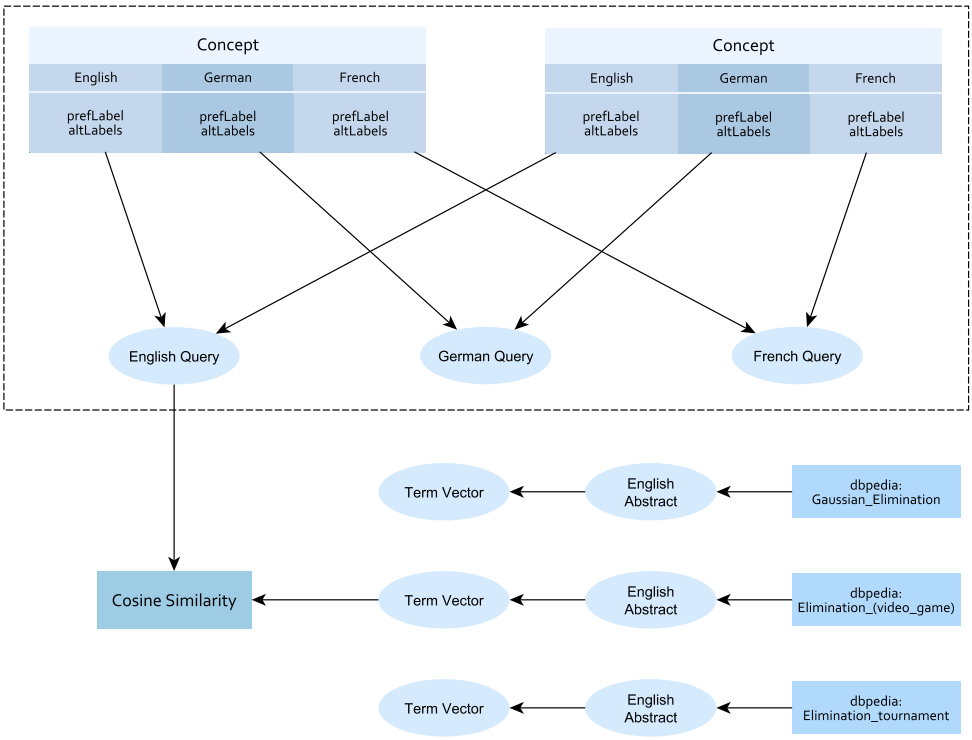


Figure 10: To emphasize the independence of the constituting concept labels, the upper part of the diagram depicts generic artifacts. For determining the correct resource for the concept *Elimination*, the term vectors of three candidates are compared to the English query.

5.4 Evaluation

	Abstracts	Full articles
Precision	92.02 %	92.64 %
Recall	69.44 %	69.9 %
F-measure	79.15 %	79.68 %

Table 5: Graph-based disambiguation followed by corpus-based approach, resulting in 1191 enriched concepts in total

	Abstracts	Full articles
Precision	67.27 %	70.9 %
Recall	85.65 %	90.28 %
F-measure	75.36 %	79.42 %

Table 6: Graph-based disambiguation followed by model-based approach, resulting in 1303 enriched concepts in total

When comparing the performance of the two-pass procedures, one can observe an opposing behavior in terms of precision and recall. Using the corpus instead of the model yields significantly higher precision since related documents, in particular in mathematics and statistics, share not only common terminology, but also similar wording and collocations. In addition, the model contains terms which can occur with similar phrasing in many different topics, resulting in more falsely classified resources. On the other hand, graph- and corpus-based disambiguation has a considerably lower recall due to quality of the corpus. Only a smaller part of the ontology concepts can be linked to some textbook index, thus for 112 ambiguous concepts no corresponding document can be obtained during disambiguation. As the index-ontology mapping depends on the focus of the textbooks, this problem could be tackled with a broader selection of books.

For graph- and corpus-based disambiguation, the choice of full article texts over abstracts for computation of cosine similarity does not result in a significant performance gain. The introductory sections obtained from the textbooks are mostly definitions and as such similar in length to Wikipedia abstracts as well as in terms of a more condensed form of language. Since the abstracts are included in full article texts, using these for comparison gives analogical matches. The larger performance gain for graph- and model-based disambiguation can be explained by the aforementioned condensed form of language. More concept labels, i.e. terms from probabilities and statistics, can be found in full articles because article bodies usually feature derivations and examples which make use of further mathematical ideas.

Due to the opposing behavior in terms of precision and recall, the difference between the F-measures of graph- and corpus-based and graph- and model-based disambiguation using full article texts is not statistically significant.

	Abstracts	Full articles
Precision	68.73 %	70.55 %
Recall	87.5 %	89.81 %
F-measure	76.99 %	79.02 %

Table 7: Graph-based disambiguation followed by corpus-based and then model-based approaches, resulting in 1303 enriched concepts in total

At first glance, the almost identical performances of the three-pass procedure consisting of graph, corpus and model and the graph- and model-based procedure seems surprising. While developing the different algorithms, the three-pass procedure has been expected to perform best because of the interplay of all different methods. However, evaluation has shown a considerable overlap between the true positives of corpus-based and model-based disambiguation. This means, the corpus-based method correctly disambiguates a larger part of the resources, which would have been also correctly disambiguated by the model-based method. The latter is then left with the ones which it classifies falsely. Therefore, the combination of corpus and model does not increase the

performance of the disambiguation algorithm.

This result indicates a class of ambiguous words which are inherently harder to disambiguate under the chosen techniques and assumptions. Evaluating further and probably more advanced approaches will be a part of future work.

6 Conclusion

This thesis has presented an automated framework for learning of a semantic, multilingual domain model. After extracting a plain SKOS model from a multilingual online glossary, its constituent concepts are mapped to resources on the linked open dataset DBpedia. These resources provide two types of information which can be used to enrich the model. Lexical information in form of abstracts results in concept definitions. The retrieval of categories and related resources introduces semantic information to the model since they represent hierarchic and associative relations. In terms of ontology learning, the framework learns taxonomic as well as non-taxonomic relations from linked open data in a domain-dependent scenario[1].

An issue with DBpedia has been shown which impedes the retrieval of multilingual information. Missing sameAs links between equivalent resources in different languages indicate misalignment of the corresponding Wikipedia articles. Although it is unclear to what extent this can be classified as inconsistent (DBpedia just accurately depicts the apparently intended misalignment on Wikipedia), such missing links constitute a considerable obstacle for cross-lingual exploration of linked data with DBpedia being a central hub of the Linked Open Data cloud[7, 8].

During the linking of model concepts to DBpedia, the problem of ambiguity arises. For polysemous query terms, disambiguation pages suggest a range of candidate resources, each representing a distinct word sense. For disambiguation of these named entities, a multi-pass algorithm combining different techniques has been presented. All three methods leverage knowledge on the specific considered domain. A graph-based approach uses the semantic relations learned during enrichment to infer domain membership of ambiguous resources based on its closed structure. Lexical domain knowledge is used by the corpus- and model-based approaches, one can further discriminate between the use of collocational and terminological information. Both methods employ the vector space model. It can be observed that the graph is able to disambiguate resources without errors just by relying on structural properties. Deciding between the corpus- and the model-based approaches on the other hand is a severe trade-off between precision and recall.

An additional flaw of DBpedia will be the main ground for future work. The English DBpedia, in contrast to the German and French, does not contain information on related resources. Consequently, many potential relations cannot be learned which significantly affects graph-based disambiguation. One could circumvent this issue by extracting the missing links from Wikipedia itself.

Furthermore, lemmatization could be incorporated during the linking of the model to

DBpedia. The model contains many verbs as labels. Since resources on DBpedia are usually stored in noun form, a considerable part of concepts cannot be linked. With proper lemmatization, the number of enriched concepts will increase.

References

[1] W. Wong, W. Liu, and M. Bennamoun, "Ontology Learning from Text: A Look back and into the Future," *ACM Computing Surveys (CSUR)*, vol. 44, no. 4, p. 20, 2012.

[2] R. Navigli and P. Velardi, "Semantic Interpretation of Terminological Strings," in *Proc. 6th Int'l Conf. Terminology and Knowledge Eng*, pp. 95–100, 2002.

[3] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to Wordnet: An On-line Lexical Database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[4] A. Oliveira, F. C. Pereira, and A. Cardoso, "Automatic Reading and Learning from Text," in *Proceedings of the International Symposium on Artificial Intelligence (ISAI)*, 2001.

[5] T. Apted and J. Kay, "Mecureo ontology and modelling tools," *International Journal of Continuing Engineering Education and Life Long Learning*, vol. 14, no. 3, pp. 191–211, 2004.

[6] T. Berners-Lee, "Linked Data - Design Issues (2006)," *http://www.w3.org/DesignIssues/LinkedData.html*, 2011.

[7] M. Schmachtenberg, C. Bizer, A. Jentzsch, and R. Cyganiak, "Linking Open Data cloud diagram," *http://lod-cloud.net/*, 2014.

[8] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, *et al.*, "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia," *Semantic Web Journal*, vol. 5, pp. 1–29, 2014.

[9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, pp. 722–735, Springer, 2007.

[10] A. Weichselbraun, G. Wohlgenannt, and A. Scharl, "Refining Non-Taxonomic Relation Labels with External Structured Data to Support Ontology Learning," *Data & Knowledge Engineering*, vol. 69, no. 8, pp. 763–778, 2010.

[11] D. B. Lenat, "Cyc: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[12] C. Matuszek, J. Cabral, M. J. Witbrock, and J. DeOliveira, "An Introduction to the Syntax and Content of Cyc," in *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pp. 44–49, 2006.

[13] T. Schandl and A. Blumauer, "Utilizing, creating and publishing Linked Open Data with the Thesaurus Management Tool PoolParty," in *OKCon*, pp. 42–48, 2010.

[14] E. Klein, B. Alex, and J. Clifford, "Bootstrapping a historical commodities lexicon with SKOS and DBpedia," *EACL 2014*, p. 13, 2014.

[15] M. A. Halliday and R. Hasan, "Cohesion in English," *English, Longman, London*, 1976.

[16] J. Morris and G. Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text," *Computational linguistics*, vol. 17, no. 1, pp. 21–48, 1991.

[17] R. Navigli, "Word Sense Disambiguation: A Survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.

[18] R. Navigli and P. Velardi, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 7, pp. 1075–1086, 2005.

[19] R. Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 4, pp. 678–692, 2010.

[20] S. Brin and L. Page, "Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Proc. Seventh Conf. World Wide Web*, pp. 107–117, 1998.

[21] A. Miles and S. Bechhofer, "SKOS Simple Knowledge Organization System Reference," *W3C Recommendation*, vol. 18, p. W3C, 2009.

[22] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 28–37, 2001.

[23] M.-R. Koivunen and E. Miller, "W3C Semantic Web Activity," *Semantic Web Kick-Off in Finland*, pp. 27–44, 2001.

[24] C. Caracciolo, A. Stellato, S. Rajbahndari, A. Morshed, G. Johannsen, Y. Jaques, and J. Keizer, "Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case," *International Journal of Metadata, Semantics and Ontologies*, vol. 7, no. 1, pp. 65–75, 2012.

[25] A. Phillips and M. Davis, "Tags for identifying languages," tech. rep., 2009.

[26] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250, ACM, 2008.

[27] E. Cabrio, J. Cojan, S. Villata, and F. Gandon, "Argumentation-based Inconsistencies Detection for Question-Answering over DBpedia.," in *NLP-DBPEDIA@ ISWC*, 2013.

[28] G. Töpper, M. Knuth, and H. Sack, "DBpedia Ontology Enrichment for Inconsistency Detection," in *Proceedings of the 8th International Conference on Semantic Systems*, pp. 33–40, ACM, 2012.

[29] D. Wienand and H. Paulheim, "Detecting Incorrect Numerical Data in DBpedia," in *The Semantic Web: Trends and Challenges*, pp. 504–518, Springer, 2014.

[30] R. C. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named Entity Disambiguation.," in *EACL*, vol. 6, pp. 9–16, 2006.

List of Figures

List of Tables