AUTOMATIC EXTRACTION OF DOMAIN-SPECIFIC ONTOLOGIES FROM TEXT RESOURCES

By
Athanasios B. Tegos

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
TECHNICAL UNIVERSITY OF CRETE
CHANIA, GREECE
MAY 2009

TECHNICAL UNIVERSITY OF CRETE
DEPARTMENT OF
ELECTRONICS AND COMPUTER ENGINEERING

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "**Automatic Extraction of Domain-Specific Ontologies from Text Resources**" by **Athanasios B. Tegos** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: May 2009

Supervisor: ______________________
Assoc. Prof. Alexandros Potamianos

Readers: ______________________
Prof. Vasilis Digalakis

____
Assoc. Prof. Euripides Petrakis

TECHNICAL UNIVERSITY OF CRETE

Date: **May 2009**

Author:	**Athanasios B. Tegos**
Title:	**Automatic Extraction of Domain-Specific Ontologies from Text Resources**
Department:	**Electronics and Computer Engineering**

Degree: **M.Sc.** Convocation: **May** Year: **2009**

Permission is herewith granted to Technical University of Crete to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

____
Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

This thesis presents our approach on automated learning of ontologies from texts semantically annotated with instances of ontologies' concepts. This approach aims to extract the ontological relational schema, which contains the semantic (taxonomic and non-taxonomic) concepts' relations, and also to learn a set of inference rules/constraints for these concepts and their relations. In comparison with other relevant approaches, the relation extraction process is not based on commonly used assumptions, that verbs typically indicate semantic relations between concepts and does not exploit lexico-syntactic patterns like Hearsts patterns, clustering methods like LSA(Latent Semantic Analysis) or any external knowledge sources like WorldNet.

Our approach is based on the assumption that concepts which are semantically related, tend to be "near" as context in a plain text. This assumption arises from the principle of coherence on linguistics. Motivated from this assumption we developed a statistical technique, which applied to documents metadata is able to discover directed semantic relation. The prior knowledge for our methodology is a text corpus annotated with instances of the concepts that we want to discover their semantic relationships. In our research we distinguish the concepts in two types, as High-Level and as Low-Level concepts according to complexity of the domain that they represent. Low-Level, are the concepts where their instances are associated with relevant text portion in the document. On the other hand, High-Level are the concepts that are "compound" in such a way that instances of these concepts are related to instances of Low-Level concepts. The extracted metadata are, the overlapping instances of the different concepts in the annotated texts. Moreover, our approach is able to discover a set of inference rules/constraints about the extracted ontology. More specifically, we propose a methodology, based on the same assumption, in order to find the type of connectivity on concepts relations. Finally, we have expounded an algorithm based on set theory that is able to discover a minimum set of rules which "describe" semantically the concepts in the level of the semantic info that their instances must contain. The proposed method was applied to corpora from two different domains, athletics and biomedical, and was evaluated against the existing manually created ontologies for these domains.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Associate Professor Alexandos Potamianos and Research Director Vangelis Karkaletsis for their valuable guidance. They provided with true generosity all the “inspiration resources” upon which this research effort is built.

I would also like to thank professors Vasilis Digalakis and Evripidis Petrakis for participating in the examining committee.

Furthermore, I would like to thanks also my colleagues from NCSR SKEL laboratory for the fruitful discussions that we had, they helped me to become familiar with research fields that are related with my work with a broader sense.

Finally I would like to thank my family, especially Vasili and Georgia.

Chapter 1

Introduction

Ontologies are widely used in Artificial Intelligence, Knowledge Engineering, Knowledge Management, Natural Language Processing and Information Retrieval. The importance of ontologies has re-emerged with the proposal of semantic web [3] by Tim Berners Lee. Originally, ontologies had a confined applicability since they are used in closed domains such as molecular biology, bioinformatics, etc to assist in knowledge management, in question answering, for information extraction, and text summarization systems. Nowadays, ontologies are the basis for the semantic web. In particular, ontologies are useful to share the common understanding of the domain between agents, to enable the reuse of knowledge, to make domain assumptions explicit, and to analyze the domain knowledge.

1.1 Ontologies

The term *ontology* comes from the Greek term *ontologia* and means “talking”(*-logia*) about about “being”(*on* ⇒ *onto-*). Ontology is a philosophical discipline which can be described *as the science of exixtence* or *the study of being.* In modern computer science parlance, one does not talk anymore about ’ontology’ as the science of existance, but as research area.

Various definitions are presented in the literature for the ontology. The following seven distinct definitions are collected and analyzed in [16] and [18] .

1. Ontology as Philosophical discipline.

2. Ontology as in informal conceptual system.

3. Ontology as a formal semantic account.

4. Ontology as a specification of conceptualization system.

5. Ontology as a representation of a conceptual system characterized by specific formal properties and only by its specific purposes.

6. Ontology as the vocabulary used by logical theory.

7. Ontology as a specification of logical theory.

Even though various definitions exists for ontology in the literature, the most appropriate motivating the ontology extraction task are by Gruber [17]. Where ontology is defined as an explicit specification of a conceptualization. According to this definition, ontology is interpreted as the formal representation of the conceptual model underlying a certain domain, describing it in a declarative fashion. As an analogy, one can describe the ontology of a domain as the relational schema of a database. The relational schema represents both the entities(concepts) and the dependency relations between the entities, whereas the ontology consists of concepts and semantic relations between the concepts.

Let's sketch the main components of an ontology, according to the previous definition. In general, an ontology of a domain consists of four major components listed below.

- **Concepts**: Concepts of a domain are an abstract or concrete entities derived from specific instances or occurrences.

- **Attributes**: Attributes are characteristics of the concepts which may or may not be concepts by themselves.

- **Taxonomy**: Taxonomy provides hierarchical semantic relations (IS-A) between the concepts. For example the concepts Vehicle and Car are related with an IS-A relation

- **Non-Taxonomic Relations**: Non-taxonomic relations specify non-hierarchical semantic relationships between the concepts. For example relations of type (has-part-some-of), like the relation between the concepts Vehicle and Wheels.

Along with the above four components, ontologies also consist of instances for each of the concepts, and inference rules of domain. This thesis emphasizes on discovering Taxonomic and Non-Taxonomic relations and also learning inference rules about concepts and relations. Detailed discussion of ontological components and existing techniques to extract each of the components are presented in chapter 2.

1.1.1 The Role of Ontologies in the Semantic Web

Ontologies possess a wide variety of applications in knowledge management, information retrieval, information extraction, question answering systems, and artificial intelligence but as mentioned above ontologies have a central role in development of the semantic web. In particular, ontologies are useful for sharing the common understanding of a domain between agents. To make the semantic web dream into reality, annotation of web pages with ontological information is necessary.

A brief description on the role of ontologies in semantic web is as follows. One of the applications of semantic web, is replacement of key word based web search with the knowledge level querying. At present search technologies, retrieve web pages arranged with efficient page ranking algorithms, consisting of key words of the user query. In this scenario, the user has to read all (or the most of) the web pages retrieved to find the answers to his query. Whereas in semantic web, each of the websites is annotated with ontologies. Hence, the whole web consists of agglomerations of domain-specific ontologies. In semantic web, the user query is analyzed at knowledge level and will be answered by performing logical inferencing using ontologies.

Considering the above notion of semantic web, various components are involved in realizing the semantic web applications. Few of them are as follows:

1. Languages for representation of ontologies.

2. Web scalable algorithms for logical inferencing.

3. Acceptability of communities(either users or businesses) for change.

4. Creation of domain specific ontologies.

As part of the language standards, various meta languages such as XML, RDF, and OWL are developed for encoding ontologies of the domain. Several algorithms and techniques for merging and querying ontologies are developed, based mainly on Description Logic. The most difficult issue is to make users accept the semantic web technology. But the most critical is the fourth component, creation of ontologies. It is required to represent websites or domain texts in terms of ontologies using one of the ontology languages. This thesis presents a novel method for automatic extraction of a domain specific ontology from annotated text resources.

1.1.2 Development of Ontologies

Even though there exists a wide variety of applications of ontologies, until nowadays, ontologies for various domains are developed manually. Some of the issues involved in the design and development of ontologies are, the requirement of expert knowledge of the domain, extensive group discussions in understanding the view point of the domain and incremental modifications to the ontology. For example, as mentioned in [36], building the initial ontology for *myGrid* project, in bioinformatics domain, took two months for an ontology expert with four years of experience in building the description logic based biomedical ontologies. In general, the construction of ontologies require the steps similar to steps involved in software development life cycle. But as in software development there are no such standards(or methods) established for the development of ontologies. Because of the lack of standards in ontology development, manual construction of ontologies is costly both in time and labor.

In computing literature, various approaches or guidelines are presented for manual construction of ontologies. Several tools such as Ontolingua [11], OilEd [1], Protege [43] and OntoEdit [44] are developed for the construction and management of ontologies. Most prominent of these are Protege and OntoEdit. The main objective of these tools is to assist the domain expert in the construction of ontologies.

To reduce the effort in design and development of ontologies, this thesis presents our proposed methodology for automatic extraction of the ontological relational schema and also discovering a set of inferred rules for the extracted ontology, from texts which are semantically annotated with instances of the ontologies' concepts. The methods presented here are domain and language[1] independent and extract each of the aforementioned components automatically.

1.2 Ontology Learning from Text

As we mentioned in the previous section 1.1.2, we propose an automatic methodol for ontology learning from text resources. Now, let's examine some of the main problems and constraints that arise from the process of ontology learning from text resources and how our proposed methodology face them.

1.2.1 What is Ontology Learning from Text?

Before presenting the problems and constraints, let's first explain the process of ontology learning and in particular the process of ontology learning from text. The term

[1] We haven't evaluated yet our proposed methodology in another language except from English but we claim that is also language independent since we do not use any syntactic or grammatic rule also our proposed methodology is based on a general assumption common in all languages.

ontology learning was originally coined by Alexander Mädche and Steffen Staab [28] and can be described as the acquisition of a domain model from data. It is historical connected to the Semantic Web, which builds on ontology models or logic formalisl restricted to decidable fragments of first-order logic, in particular description logics [42]. Thus, the domain models to be learned are also restricted in their complexity and expressivity.

Obviously, ontology learning needs input data from which to learn the concepts relevant for a given domain, their definitions as well as the relations holding between them. One crucial requirement is thus that the input data is representative for the domain one aims to learn an ontology for. In case ontology learning is performed on the basis of unstructured textual resources, we will speak of ontology learning from text. More specifically, ontology learning from textual resources can be regarded to some extent as a process of reverse engineering. Because the author of a certain text or document has a world or domain model in mind which he shares to some extent with other authors writing texts about the same domain. Thus, during the process of ontology learning from textual resources we try to reconstruct this implicit domain model of the author or even of the model shared by different authors.

1.2.2 Constraints on Ontology Learning from Text

You can perceive from the previous section 1.2.1 that the process of reconstructing the implicit domain model, ontology learning, from unstructured text is a very difficult process. Because of constraints both in text processing and knowledge acquisition.

Natural language texts are not only unstructured but also ambiguous in word meanings and usage. Because of the unstructuredness and ambiguousness, it is difficult to perform semantic analysis of natural language texts. Methods for ontology extraction methods must address the following issues with respect to text processing.

Unstructured Text Even though the natural language processing research has its origins from the early 50s, because of the unstructuredness of the text, there exists no fixed schemata for interpreting the natural language statements. Hence it is very difficult to convert unstructured texts into a structured representation which is required for computer processing systems.

Ambiguity in English Text In addition to the unstructuredness, natural language text is also ambiguous. The meaning of a word varies based on the context in which it occurs. In natural language processing, context in which a word can occur is defined as the "*sense*" of the word. In English language, most of the nouns consists of more than one sense. For example, the word "*form*" has sixteen different senses. A word not only consists of multiple senses but also appears in multiple parts of speech. For example, the word "*like*" can occur in eight distinct parts of speech as shown below.

- verb "Fruit flies *like* a banana."
- noun "We may never see its *like* again."
- adjective "People of *like* tastes agree."
- adverb "The rate is more *like* 12 percent."
- preposition "Time ies *like* an arrow."
- conjunction "They acted *like* they were scared."
- interjection "*Like*, man, that was far out."
- verbal auxiliary "So loud I *like* to fell out of bed."

Lack of Closed Domain of Lexical Categories The possible lists of pronouns, prepositions, conjunctions, and interjections are fixed. These four parts of speech are called as functional or closed categories. Elements of the remaining parts of speech such as nouns, verbs, adjectives, and adverbs are not fixed. These are called lexical or open categories. New words are added to the English dictionaries from other languages or some other sources. Variations of the nouns are used as adjectives, verbs, and adverbs. Because of the lack of closed sets for lexical categories, it is very difficult to identify the validity of the extracted terms automatically.

Noisy Text When large amounts of text is collected for processing, there exists a very high possibility for the presence of noise in the collected text. It is difficult to identify and filter such noisy text without knowing the content of the whole text. For example, texts may contain analogies and metaphors which are not relevant to the domain text.

Lack of Standards in Text Processing In general, text documents written by various authors convey different perspectives of the domain. Hence all the documents may not represent the same view of the domain. Even all documents may belong to a single domain, some of them may support and others may oppose a view point. It is very difficult to identify the collections of text which represent the single view of the domain. No standards are established in identifying the perspective (Example: supports or opposes an idea) of a document.

Moreover, except from the limitations in text processing there is also constraits in knowledge acquisition. Some of them are described below.

Lack of Fully Automatic Methods for Knowledge Acquisition Because of the unstructuredness and ambiguity in texts, no fixed procedures exist for the analysis of text. Many of the existing techniques for knowledge acquisition from texts are domain dependent and based on supervised learning methods. Supervised learning methods require large amounts of training data for each of the domains. Since the ontology represents the knowledge of the whole domain, it is difficult to have large amounts of the data for training to build the ontology and additional data for ontology extraction. Ideally, the techniques for ontology extraction should be domain independent and should not rely on large amounts of training data.

Lack of Techniques for Coverage of Whole Texts Most of the current approaches in the literature are developed with the aim of building or extending thesaurus from the texts. Existing approaches are based on the word frequencies, syntactic-patterns, grammatical-rules and heuristics-rules according to the training corpus. These approaches cover only terms or sentences which satisfy the above constraints. The remaining text is ignored. But most of the ignored text also contains useful knowledge about the domain. It is desirable to develop the techniques which covers the most of text possible.

1.3 Our Proposed Methodology

Because of the several constraints mentioned above, automatic ontology extraction is a difficult task. Motivated from the constraints mentioned before and trying to develop an automatic and domain independent method, we have expounded a statistical methodology, applied on metadata extracted from the corpus, for ontology learning. We adopted a statistical approach in order to face both constraints arisen from text processing as well as to overcome the lack of domain independent techniques for knowledge acquisition from text. Moreover, our proposed methodology does not based on commonly used approaches for knowledge acquisition like syntactic-patterns, grammatical-rules or lexical knowledge bases. Our method is based on a principle of linguistics and more specifically on the principle of coherence on linguistics [34].

As we mentioned above 1.1, domain ontologies consist of concepts, semantic relations among these concepts, and a set of inference rules. Thus, the process of ontology learning from text includes three core subtasks: learning of the concepts that will constitute the ontology, learning of the semantic relations among these concepts and finally, learning of a set of inference rules.

This thesis presents our approach on automated learning of ontologies from texts that are semantically annotated with instances of ontology concepts. This approach aims to discover the ontological relational schema, which contains the semantic relations (taxonomic and non-taxonomic) between the concepts that have been annotated, as well as to discover a set of inference rules for these concepts and their relations.

In our approach we distinguish the concepts in two types, as High-Level and as Low-Level concepts depending on complexity of the domain that they represent. More specifically, we categorize as Low-Level concepts, the concepts where their instances are associated with relevant text portion in the document. On the other hand, we categorize as High-Level the concepts that are "compound" in such a way that instances of these concepts are related to instances of Low-Level concepts. According to this categorization, the concepts *"Name"* or *"Nationality"* of a person are Low-Level concepts, because their instances are associated with relevant text portion in the document and are directly identifiable, to the contrary of the concept *"Person"* where its instances are related to instances of concepts *"Name"* and *"Nationality"* and are not directly identifiable in a text document. A more detailed explanation, about the discrimination between High-Level and Low-Level concepts, is presented in section 3.4. In comparison with other relevant approaches, we focus on the discovery of semantic relations between High-Level concepts, but we also show the applicability of the proposed approach to Low-Level concepts.

The discovery process of semantic relations is not based on commonly used assumptions that are used in the literature, that verbs typically indicate semantic relations between concepts or does not exploit syntactic-patterns or clustering methods or any external knowledge-base like WorldNet. Our approach is based on the assumption that concepts which are semantically related, tend to be "near" as context in a plain text. This assumption arises from the principle of coherence on linguistics. Based on this assumption, statistical methods are applied to metadata extracted from the annotated texts, to discover semantic directed relations between concepts. Moreover, we propose a methodology, based on the same assumption, in order to find the type of connectivity on concepts' relations. Finally, we have expounded an algorithm based on set theory that is able to discover a minimum set of rules which define semantically the least info that must exist for the representation of a High-Level concept's instance.

In comparison with other relevant approaches, we propose a methodology that is able to automatically extract, from a corpus related with a domain, an ontology related with this domain. Where the extracted ontology is constituted by any type and any number of concepts, furthermore our approach is also able to discover inferred knowledge about the concepts and the relations of the extracted ontology.

The remainder of this thesis is organized as follows. Chapter 2 explains each of the sub tasks and provides a detailed discussion on the existing approaches and their shortcomings. The proposed method for the discovery of semantic relations between High-Level concepts but also between Low-Level concepts are presented in chapter 3. Similarly, chapter 4 presents the methods developed for finding the inference rules for the concepts and the discovered relations. The experimental results and the evaluation of the proposed method for two different domain ontologies are presented in chapter 5. Finally, conclusions and future directions are presented in chapter 6.

1.4 Summary

Ontology of a domain consists of concepts, semantic (taxonomic and non-taxonomic) relations between the concepts and a set of inference rules. Ontologies are widely used in information retrieval, artificial intelligence, and intelligent information integration tasks. The importance of ontologies has re-emerged with the proposal of semantic web. Even though ontologies posses a variety of applications, as of now ontologies are developed manually. But manual construction of ontologies is costly both in time and labor. To reduce the effort in manual construction of ontologies, this thesis presents a novel method on automated learning of ontologies from texts that are semantically annotated with instances of ontology concepts. This approach aims to discover the ontological relational schema, which contains the semantic relations (taxonomic and non-taxonomic) between the concepts that are annotated, as well as to discover a set of inference rules for these concepts and their relations.

Our approach is based on the assumption that concepts which are semantically related, tend to be “near” as context in a plain text. Statistical techniques are applied to metadata extracted from the annotated texts, to discover semantic relations among the annotated concepts as well as to find the type of connectivity on concepts’ relations. Moreover, we propose an algorithm based on set theory that is able discover a minimum set of rules which define semantically the High-Level concepts. The proposed method was applied to corpora from two different domains, athletics and biomedical, and was evaluated against the existing manually created ontologies for these domains.

The next chapter reviews the literature of methods that have been proposed for extraction of each of the aforementioned sub tasks.

Chapter 2

Literature Review

As mentioned in the previous chapter, with the advent of semantic web, many ontology engineering projects has started. But most of the projects are still in at their infancy. Some of those are Text-to-Onto [29] and Hasti [40]. Text-to-Onto is a part of the KAON(KArlsruhe ONtology) Tools for ontology management. It supports semi-automatic creation of ontologies using text mining algorithms. Currently, the tool includes concept extraction and concept association extraction algorithms. Text-to-Onto is embedded in in the ontology editing tool OntoEdit [44]. OntoEdit allows to browse and edit the existing ontological concepts. Text-to-Onto extracts the conceptual structures using the term frequencies from text. Concept hierarchy is extracted using hierarchical clustering algorithms and non-taxonomic relations are extracted using association rule mining algorithm. Text-to-Onto tool requires users verification at each stage of the ontology extraction process. For example, concepts extracted using frequency counting need to be verified before finding the relations between the concepts. Also, it requires manual labeling of internal nodes of the hierarchical clusters to find the taxonomic relations.

Similar to Text-to-Onto, Hasti is another tool developed extracting ontologies. Hasti is developed for processing Persian texts. Hasti operates in both cooperative and unsupervised modes. In cooperative mode, the user decides the selection or rejection at each stage of the process. For example, the user has to select the concepts from the candidate ones. In the unsupervised mode the system automatically selects each of the components of the ontology. In an overview, Hasti, initially, accepts a few top-level concepts, taxonomic and non-taxonomic relations as kernel elements, and extends initial seeds by adding more concepts. These kernel elements are linguistically motivated concepts like object, action, property, and etc. In Hasti, to extract the candidate concepts, a set of rules are defined to identify the structural sentences. A set of sentences matching one of the rules are considered as candidates. From the candidate sentences, candidate concepts are extracted by identifying nouns using predefined structures. To find the taxonomic and non-taxonomic relations, both hierarchical and non-hierarchical clustering algorithms are used. In addition to clustering algorithms, Hasti uses predefined semantic templates to extract the knowledge from the candidate sentences. Hasti is an ongoing project and also does not report any new methods on identification of relations.

Along the lines of Text-to-Onto and Hasti, several other organizations have started various projects for ontology extraction such as ASIUM [12] and FFCA [32]. ASIUM learns semantic relations by clustering the nouns based on their occurrence with the verbs. In ASIUM each of the clusters of nouns is presented to the user for labeling. FFCA incorporates fuzzy logic into formal concept analysis for learning ontologies. In FFCA, concepts are extracted based on fuzzy membership value associated with each context. Conceptual relations between the concepts are obtained using fuzzy conceptual clustering algorithms.

Even though there is a lack of much work on extraction of full scale ontology extraction systems, considerable research has been done in the extraction of individual components of the ontologies. The following sections describe the existing approaches in the acquisition of each of the components of the ontology. More specifically, approaches and methods for semantic relation extraction and inference rules acquisition.

2.1 Taxonomic Relation Extraction

For automatic construction of ontologies, we need techniques for automatic extraction of both hierarchical and non-hierarchical relations. Existing methods for semantic relations extraction are presented in two separate sections as hierarchical relations extraction and non-hierarchical relations extraction.

In the literature, hierarchical relations among the concepts are also called taxonomic relations or simply taxonomy. Existing techniques for finding taxonomic relations can be classified as pattern based, clustering based approaches, and combination of both. In pattern based approaches, the user defines a set of predefined lexico-syntactic patterns. Domain text is verified against the patterns to obtain the instances of taxonomic relations. In clustering based approaches, hierarchical clustering algorithms are used for finding the taxonomic relations between the concepts. And heuristics are used for labeling the internal nodes in the clusters. In the combined approaches, internal nodes are labeled using the instances extracted using lexico-syntactic patterns.

One of the early works for finding taxonomic relations based on lexico-syntactic

No	Syntactic Pattern	Hyponym Relation
1.	NP_0 such as $\{NP_1, NP_2, \ldots, (and \mid or)\}\ NP_n$	hyponym (NP_i, NP_0)
2.	such NP_0 as $\{NP_i, *(or \mid and)\}\ NP_n$	hyponym (NP_i, NP_0)
3.	$NP_1\ \{, NP_i\} *\ \{,\}\ (or \mid and)$ other NP_{n+1}	hyponym (NP_i, NP_{n+1})
4.	$NP_0\ \{,\}\ (including \mid especially)\ \{NP_i\} *\ \{or \mid and\}\ NP_n$	hyponym (NP_i, NP_0)

Table 2.1: Hearsts Patterns for Taxonomic Relation Extraction

patterns is presented by Hearst [19]. Hearst's patterns and their corresponding hyponym relations are shown in Table 2.1. Hearst's procedure to identify the hyponym relations is as follows. Extract the sentences which satisfy any of the patterns listed in Table 2.1. For each sentence, identify the noun phrases which satisfy corresponding NP in the pattern. Label the relation among the noun phrases using the corresponding hyponym relation of the pattern. For example, the sentence,

"The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string".

satisfies the pattern 1 in Table 2.1. Here, NP_0 corresponds to "$bowlute''$ and NP_n corresponds to "$Bambarandang''$. Hence, the hyponym relation extracted is:

$hyponym($ "$Bambarandang''$, "$bowlute''$ $)$

It is quite intuitive that the authors mention such sentences as illustrations of the unknown terms meaning identification. Further more, Hearst presented a simple heuristic to extract the instances of additional relations as follows. Select a set of pairs of terms which satisfy the target semantic relation for bootstrapping. Extract the sentences which consist of pairs of terms. From the extracted sentences, identify the commonalities and hypothesize the common structures that yield patterns of target relation. Even though the above heuristic seems to work, Hearst mentioned that they didn't get much success in extracting the meronym (i.e. part-whole) relations.

Taxonomic knowledge acquisition technique presented in [21] is similar to Hearst's approach. She presents mainly two lexico-syntactic patterns for taxonomy extraction. One of the patterns is same as pattern 2 in Table 2.1. Another pattern consists of, if a pair of terms connected by the verb *like.* But, according to [21], large number of pairs extracted with the *like* pattern are spurious. Simple heuristic rules are proposed to reduce the spurious relations and to identify the concept boundaries. Detailed experimentation of the patterns is performed on the *Time Magazine* corpus.

A framework for acquisition of hypernym links among multi-word terms using single-word candidates is presented in [30]. This system is built based on the previous work described in [19]. It provides a classifier for the purpose of discovering new lexico-syntactic patterns through corpus exploration for the given semantic relation. More specifically, the system is a combination of *Promethee*, a tool for structuring the relationships among the single-word terms, *ACABIT* [9], a tool for acquisition of multi-word terms, and *FASTR* [22], a tool for term variant recognition of the candidate terms. Finally, the system inherits the relations between the single word terms to the corresponding multi-word variants. The *Promethee* system extracts lexico-syntactic patterns for the given semantic relation using a set of terms which satisfy the relation. In summary, *Promothee* collects sentences from the corpus and determine the patterns of the sentences in which the above seed terms are present. Additional sentences satisfying the patterns are extracted to find more instances of the semantic relation. The *Promethee* system is experimented with three (hypernym, merge, produce) relations. Similar to techniques, the *Promethee* also extracts relations between the terms which occur in the same sentence only. To find the relations between the terms across different sentences, the system tries to identify the variations of the terms for which the relations are already determined, then the same relation assigns to the variants. For example, if the relation between *fruit* and *apple* is known then the relation between multi-word variants like *fruit juice* and *apple juice* is also labeled as same. The *FASTR* extracts multi-word terms using syntactic, morpho-syntactic, ans semantic categories of variations. For each of the categories, various rules are defined to identify the multi-word terms. Semantic relations among the multi-word terms, with reference to semantic relations among their constituent words, are labeled if the following three constraints are satisfied.

Semantic Constraint Two multi-word terms w_1w_2 and $w_1^{'}w_2^{'}$ are semantic variants of each other if the following three constraints are satisfied.

1. Some type of semantic relation S holds between w_1 and $w_1^{'}$ and/or between w_2 and $w_2^{'}$.

2. w_1 and $w_1^{'}$ are head words and w_2 and $w_2^{'}$ are arguments with similar thematic roles.

3. w_1w_2 and $w_1^{'}w_2^{'}$ share the same type S of semantic relation.

The above technique for finding semantic relations among multi-word terms also provides the opportunity to cluster the semantically related words. This technique provides the opportunity to increase the recall in terms of the number of relations extracted and also the coverage of the terms. It is able to find the relations among multi-word terms which occur in two different sentences. But to label such relations, the relation between their constituents should be known by some other means. The expert's intervention is required to validate the patterns identified by the *Promethee* using the seeds. Extraction of the seeds from the knowledge base for the given semantic relation also requires human involvement. Taxonomic relations among the terms which does not follow the preselected patterns are not retrieved using the above mentioned system.

Even though the above patterns retrieve valid taxonomic relations, these patterns extract hyponym relations between the concepts which occur only in the predefined patterns. According to the results presented in the corresponding works, the number of hyponym relations extracted comparing the size of the corpus is very small. These approaches may have high precision because most of the extracted relations are valid but produce a low recall because of the occurrence of few such patterns in domain text. Further more, corpus used in these experiments does not belong to a fixed domain. The disadvantage of the pattern based approaches is that these approaches find pairs of nouns which hold taxonomic relations rather find the relation between the given concepts. Hence pattern based approaches may be suitable for extending thesaurus but not for ontology acquisition.

To build a hypernym-labeled tree from text, Caraballo [4] presented a technique based on cosine similarities using bottom-up clustering. The input to the technique is a set of nouns which are separated by conjunctions or appeared as appositives. Using the frequency of occurrence of each word along with the other words as the criteria, similarity of the words is determined by the cosine metric. Two nouns which are highly similar are grouped by giving them the common parent. The process is repeated until a single parent is found for all the nouns. Similarities among the internal nodes are determined using the weighted measure of the similarities of their leaves. Labels for the internal nodes are determined using their leaves and the Hearst's patterns. Each leaf maintains a vector of hypernyms extracted using the patterns. For each internal node of the tree, he constructs a vector of hypernyms using the hypernyms of the children. Internal nodes are labeled with the hypernym which has maximum count. For each internal node, the author suggested assigning the best, second-best, and third-best hypernyms based on their occurrence count. Also, Caraballo suggested a simple heuristic to reduce the size of the tree by eliminating the unlabeled internal nodes.

Though this technique is quite straightforward and simple, it also depends on the Hearst's patterns for labeling the hypernyms. Due to this, as the author mentioned, large number of nodes are unlabeled. Another constraint is it considers only terms with single word which occurs in the specific contexts. These words may describe only a subset of the domain. This method is also experimented on domain independent corpus.

Snow et al [41] proposed a supervised learning technique using dependency paths as features to find the syntactic patterns for hypernym relation extraction from text. The dependency paths are generated using parse trees. The training set for this approach is pairs of terms (w_i, w_j) which occur in a sentence. The pair of terms are classified as valid hyponym/hypernyms if both of them are in hypernym relation according to the WordNet[1] with the most frequent sense. The patterns for hypernym relation are discovered from the dependency paths in parse trees which occur in at least five unique hypernym/hyponym pairs in the corpus. This technique also restricts the hypernym relations between the terms in the same sentence only.

Similar to the above techniques for taxonomic relation extraction, in [5], the authors used the Latent Semantic Analysis(LSA) [47] to eliminate the invalid hyponym/hypernym pairs, and used the coordination information to improve the recall. Other related works for the taxonomy extraction are [23], [35] and [13]. But the techniques presented in [23] and [35] are specific to the medical domain text, and thus domain specific compositional terms can be exploited. In [13], hierarchies are found using document collection subsumption rule i.e. Hyponym relation between w_1 and w_2 is based on the relative frequencies that the number of documents contains both w_1 and w_2 versus number of documents in which w_2 alone is present.

Even though there exists an extensive collection of literature on taxonomy extraction, none of the presented techniques assume that all documents belongs to a single domain. As mentioned before, existing methods extract the taxonomic relations from text by identifying instance of the patterns or nouns occurring in pre-specified positions. To the best of our knowledge, none of the methods find the taxonomic relations between the given set of concepts using the text in which they occurred.

2.2 Non-Taxonomic Relation Extraction

The other major component of semantic relations extraction, during the process of ontology learning from textual resources, is the discovery of non-hierarchical semantic relations. Existing methods on extracting non-taxonomic semantic relations from texts can be classified into the following two categories.

- Approaches for extraction of concept pairs which are related with a given relationship label.

[1] A lexical database for the English language, http://wordnet.princeton.edu/

- Approaches for discovering of relationships between the concepts in a given set of concepts.

The first category of the approaches for relation extraction task are techniques for finding the concept pairs, such that their instances are related with a pre-specified semantic relationship. Some of the existing works which follow the above mentioned approach are [2], [15], [14], and [45]. Among the existing works listed above, [2] and [15] finds the noun pairs which are related with a *part-whole*[2] semantic relations. [14] presents the patterns for identification of concept pairs which are related with a *cause-effect* semantic relationship. [14]'s technique learns the semantic patterns for a given semantic relation. Learned patterns are used to find concept pairs which are related with the same semantic relationship.

Berland et al[2], presented a pattern based technique to extract the parts of the components from large corpora. To extract the patterns for *part-whole* relation, the authors used the pair *("basement", "building")* which hold the specified semantic relation and extracted all the sentences which consists of the pair. From these sentences, a set of patterns are extracted. After the manual evaluation, the number of patterns extracted are reduced to two. To extract additional pairs, for a given word, all the sentences which satisfy any of the two selected patterns are extracted. From each sentence, the noun phrase which is in the part position is extracted. All the extracted parts are ordered by the likelihood that they are true parts according to the sigdiff(significant-difference) metric. The metric is based on the idea that for a given *whole* W and *part* P, how far apart can we be sure the distributions $P(W|P)$ and $P(W)$ at the given significance level, say .05 or .01. The authors tested the above technique for six different part words for each of the whole words. After the human evaluation by six different subjects, the authors claim that the presented technique results in 55% accuracy for the top 50 words, as ranked by the system. As the author mentioned, this technique relies on very large corpus (100,000,000 words). This technique requires to provide the terms which satisfy the given semantic relation, in order to identify the patterns.

Similar to the aforementioned work, Girju et al[15] described a technique to learn semantic constraints for finding the *part-whole* relations. Though the authors didnt mention it explicitly, this technique is an extension of the work in Berland et al[2]. Here the authors able to extract three patterns shown in Table 2.2 by analyzing the TREC-9 corpus.

To identify valid *part-whole* pairs from sentences which satisfy one of the patterns in Table 2.2, the authors proposed a supervised learning technique using $C4.5$ decision tree algorithm [33] for learning semantic constraints. The attributes representing each

[2] A *part-whole* relationship indicates that one or more of one concepts is part of another concept.

No	Lexical Pattern
1.	NP_1 of NP_2
2.	NP_1 's NP_2
3.	NP_1 Verb NP_2

Table 2.2: Lexical Patterns for Part-Whole Relations

candidate pair are, WordNet class and sense number of the part and whole terms. Each noun pair is classified as whether its constituents hold a valid *part-whole* relation or not. The authors extracted 34,609 sentences as positive examples and 46,971 as negative examples. For both *part-NP* and *whole-NP* in each example, the authors assigned semantic class (WordNet class) and sense number manually. From these examples, the authors filtered out ambiguous instances by assigning more specific WordNet classes. Specialization process is repeated until the ambiguity in the input data is resolved. The $C4.5$ algorithm is applied to unambiguous examples to learn the semantic class pairs which indicate valid *part-whole* relation. The rules learned using $C4.5$ algorithm are considered as the constraints to be satisfied for any two *NPs* in order to satisfy the *part-whole* relation. Here, the authors put enormous effort in assigning the class and sense number for each pair of the *NPs* in the sentences manually. The learned semantic constraints can be used to filter some of the irrelevant noun pairs. To apply the rules learned for a noun pair, it is required to find, for each noun, the taxonomy path from noun to a top class in the WordNet. These rules are not useful for noun pairs whose constituents are not listed in the WordNet. Even for a noun pair whose both of the nouns are present in the WordNet it is required to identify their sense correctly to able use the learned rules.

In Girju and Moldovan[14], a semi-automatic technique for extraction of lexico-syntactic patterns for *cause-effect* relation is presented. This technique also relies on large corpus and the WordNet. The algorithm primarily consists of two steps. In the first step, the algorithm selects a set of pairs of noun phrases which hold the *cause-effect* relation from the WordNet. Extract the sentences which consists of the selected noun phrases and are of the of the form $\left(NP_1 \;\; verb|verb - expression \;\; NP_2\right)$ from corpus. Filter the nouns such that each of the nouns corresponds to NP_2 has to be one of the *human action*, *phenomenon*, *state*, *psychological feature* and *event* WordNet classes. The nouns corresponds to NP_1 must be subclasses of *causal agent.* The *Verb—Verb-expression* must have few number of senses and highly frequent. The causal relationship extracted using the above patterns, is validated and assigned a rank (between 1 and 4) to indicate its strength. A simple algorithm based on the WordNet classes, frequency, and ambiguity of the verbs is proposed to rank each relationship. Using the causation verbs extracted from the above approach, 50 sentences for each verb and thus around 3000 sentences extracted, on which the 1321 sentences of them are in the $\left(NP_1 \ \ verb \ \ NP_2\right)$ pattern. From these sentences, the system extracted 230 relations as valid with one of the four ranks. This approach is quite general, domain independent, and is not dependent on the hand coded patterns. But this technique requires valid instances of the causal relationship and also makes use of the external lexical knowledge base(i.e.WordNet). The number of relations extracted from the large corpus (3GB of news articles) is very few (230 relations).

We believe the presented techniques might be useful in extending the thesaurus or lexical knowledge bases. But these techniques might not be suitable for extracting ontology relations between the concepts, because of the following reasons. The first is that the extracted ontology concepts may not present in the identified patterns. The other reason, is the nouns presented in sentences which satisfy the patterns, might not have been considered as valid concepts of the domain. Further more, using these techniques, considering the amount of input text processed, only a very few pairs are identified.

The last category of approaches, for identification of non-taxonomic semantic relations, is finding the candidate concept pairs and labeling the extracted relationship between their instances. Techniques in this category need to identify the existence of a relationship between concepts in a concept pair and also to label the relationship appropriately. In [24], the authors presented a simple heuristic based on the conditional probability to label the relations between the concepts using verbs. The relation labeling technique is based on the hypothesis that predicate of a semantic relation can be characterized by the verbs frequently occurring in the neighborhood of pairs of concepts associated with it. Each triple, pair of concepts and the verb nearby $(C_1, \ C_2, \ V)$, is treated as a transaction. For a given transaction, if its frequency of occurrence is greater than the expected frequency then the verb(V) is considered as a candidate to label the relation between the concepts. All the verbs which occur above the expected frequency along with the concepts are considered as candidates for relations among the concepts. The triple $(C_1, \ C_2, \ V)$ is valid, if and only if both the concepts C_1, C_2 occur within n (experimentally $n = 8$) words from V. In the experiments, TAP knowledge base[3] is used to identify the classes of the named entities. TAP consists of a large repository of lexical entries such as proper names of places, companies, people and the like. The technique for automatic identification of the concept for a given instance is a research question. Another major drawback of this approach is the fact that it is not able to identify the direction of the relation $(C_1 \rightarrow C_2 \ \ or \ \ C_1 \leftarrow C_2)$. Also, it is not able to label the relations between the concepts whose lexical entries are connected by prepositions or conjunctions. Further more, this technique does not address the issue

[3] http://www-ksl.stanford.edu/projects/TAP/

of finding the relations among the concepts which does not occur in the same sentence. As the author mentioned the results are not impressive due to the following reasons: richness and relevance of the concept taxonomy, richness and relevance of the lexicon, style of the underlying text, performances of the PoS tagger.

Another related work comes under the second category for semantic relation extraction is presented in [7]. It is based on the x^2 test. The technique works as follows. Each occurrence of instances of the concepts in the domain text are replaced with the corresponding concepts. From the modified text, select the sentences which satisfy the pre-specified patterns. The dependency patterns are extracted for each of the sentences. For a given pair of concepts and a dependency pattern, if the occurrence of concepts as fillers of the pattern is greater than the expected frequency then the relation between the concepts is labeled with the name of the dependency pattern. The x^2 test at 95% confidence interval is used to test the hypothesis. This technique finds the relations between concepts in the same sentence only. This work is specific to the molecular biology domain. Portability to other domains is a question because structural patterns of terms varies with the domain.

Along with the above techniques, other techniques in the literature for finding the relations between the concepts are [38] and [12]. Similar to Ciaramita et al[7], Schutz and Buitelaar[38]'s work also extracts the concept pairs presented in dependency relations and use the x^2 test to verify the statistical significance on the togetherness of the concepts. Faure and Nedellec[12]'s work learns semantic relations from sentences occurring in pre-specified patterns. Nouns occurring in the pre-specified positions(subject or object) for a given verb are clustered. Each of the clusters are manually labeled with the representative concept. The verb with which a cluster is formed is considered as the label for relationship between the concepts. The main constraint of this method is it requires manual labeling of clusters with concept names in finding the relationships. Among the two different categories of methods presented, the last category of methods are more essential for finding the non-taxonimic relations between concepts.

Finally, another major part of related work for semantics extraction from textual resources is by using spacial proximity for discovering semantic similarity. The computational model of semantics is refered as *word-space model* by Hinrich Schütze [39]. A model that measures the semantic relationship between words is defined with respect to the vocabulary which forms a high-dimensional space whereas each word can be considered as one dimension. The word-space model reflects a spatial representation of word meaning. The key idea of this model is that semantic similarity can be represented as proximity in n-dimensional space, where n is the cardinality of vocabulary set [37]. Spatial proximity between words as a representation of their semantic similarity seems to be very intuitive and naturally derived with respect to the way that human conceptualize similarities. This *geometric metaphor of meaning* has been pointed out by the work of Lackoff and Johnson [25, 26]. They state that metaphors form the raw base of abstract conceptualization. Also, they argue that these metaphors are used by human mind for reasoning about abstract and complex phenomena, such as natural language and semantics. This physical tendency of human mind places the conceptual locations of words with similar meaning to be "near" each other, while the dissimilar words are placed "far apart. Of course, a sole word in a high-dimensional space gives no additional information for deeper understanding off the word. The space must be populated with other words in order to apply the proximity as an indicator of similarity. The geometric metaphor of meaning conceptualize the words as locations in a word-space and the similarity is considered as the proximity between the locations [37].

Considering the various constraints mentioned above with the existing approaches, we have developed a statistical approach that is able, without any training or other heuristics/supervision techniques, to discover semantic relations, both taxonomic and non-taxonomic, among the concepts that have been annotated in a corpus. Our proposed approach is domain independent, since it does not use pre-defined patterns and does not depended on the type or number of concepts. Moreover, is based on a very general assumption arises from linguistics. Detailed discussion of our approaches is presented in chapter 3.

2.3 Inferred Rule Acquisition

The last component, in the process of ontology learning from textual resources, is the task of discovering a set of inference rules about the domain of the extracted ontology. These rules are necessary because they describe, in a machine readable way, the implicit knowledge of domain as a set of rules or constraints on concepts and relations. The task of discovering a set of inference rules, is the least addressed aspect of ontology learning. The community did not pay yet much attention in this research field, which in our opinion is equally important because it is necessary in order to make the semantic web dream into reality. Due to the lack of extensive methods in the literature we will mention here the most important, according to our judgment.

The idea of deriving inference rules from text has been pursued in Lin and Pantel[8] and it is aimed at discovering paraphrases. They propose an unsupervised method for discovering inference rules from text, such as *"X" is author of "Y"* $\approx$ *"X" wrote "Y"*, *"X" solved "Y"* $\approx$ *"X" found a solution to "Y"*. They propose an algorithm (DIRT) that is based on an extended version of Harris' Distributional Hypothesis, which states that words that occurred in the same contexts tend to be similar. Instead of using this hypothesis on words, they apply it to paths in the dependency trees of a parsed corpus. In their approach text is parsed into paths, where each path corresponds to predicate argument relations and rules are derived by computing similarity between paths. Essentially, if two paths tend to link the same sets of words, they hypothesize that their meanings are similar. The rules in this case constitutes an association between similar paths.

The field of inductive logic programming (ILP) is also relevant to this problem. Liakata et al[27] propose a methodology, based on ILP, of automatically learning domain theories from parsed corpora of sentences from the relevant domain and using weighted finate state automation (FSA) techniques for the graphical representation of such theory. By a "domain theory" they mean a collection of rules which capture what commonly happens (or does not happen) in some domain of interest. Using WARMR, an ILP system that learn generalizations and correlations first order logic predicates, parse the input sentences into a list of frequently associated predicates, found in the flat quasi-logical forms of the input sentences. Afterwards, represents each of the extracted predicates using weighted FSAs and applying minimization and determination algorithms reduce the large set of overlapping clauses and then using the frequency information given by WARMR, calculate the weights on transitions.

Another major task in the field of ontology learning which, as far as we know, is not addressed yet is the problem of finding the type of connectivity between the instances of two semantic related concepts. We believe that this information is very important for many reasons, such as, logical inference on the ontology, discovering new instances of the related concepts for population of the ontology, etc. In this thesis we present an automatic statistical method that is able to discover the type of connectivity among the instances of the related concepts. Moreover, we have developed an algorithm based on set theory that is able to discover a minimum set of rules which define semantically the least info that must exist for the representation of a concepts instance.

2.4 Summary

Because of the various constraints in natural language processing and knowledge acquisition, extraction of domain-specific ontologies from textual resources is a diffcult task. Even though several projects are started for automatic construction of ontologies, most of them are still in at their infancy. Some of those are Text-to-Onto and Hasti. At this point, both Text-to-Onto and Hasti find conceptual terms using term frequencies and concepts association using clustering algorithms. Even though there exists a lack of research on the automatic extraction of full-scale ontologies, considerable attention has been focused on the extraction of its individual components.

In applied natural language processing, finding taxonomic relations from text is widely investigated. Existing techniques for taxonomy extraction can be classified into pattern based, clustering based, and combination of both. Most of the pattern based approaches rely on the Hearst's patterns listed in Table 2.1. In pattern based approaches, domain text is verified against the pre-specified patterns to find the instances of taxonomic relations. Since only a few sentences satisfy the pre-specified patterns, recall of these methods will be poor. In addition, pattern based approaches find taxonomic relations between nouns occurred in pre-specified patterns rather than between the concepts already identified. In clustering based approaches, hierarchical clustering algorithms are used for finding taxonomic relations between the concepts. The main difficulty with the clustering based approaches is finding labels for the internal nodes. Some of the techniques are developed for labeling internal nodes using results of the pattern based approaches.

The other major component of semantic relations extraction in automatic ontology learning, is the discovery of non-taxonomic semantic relations. Existing methods for discovering non-taxonomic semantic relations can be classified in two categories as finding concept pairs which appear in the pre-specified relation and the other is finding semantic relations between the concepts in a given set of concepts. The first category techniques, find concept pairs which are semantically related with a pre-specified relation, namely part-whole, cause-effect, etc. The last category of approaches discover semantic relations between the concepts based on the dependency patterns and statistical techniques. Main constraint of these approaches is that semantic relations between the concepts occurred in prepositional phrases and appositives, are not considered.

The least addressed aspect of ontology learning is the task of discovering a set of inference rules about the domain of the extracted ontology. An initial attempt to formulate the problem is presented by Lin and Pantel, where they presented an unsupervised method for discovering paraphrases in text corpora. The field of inductive logic programming (ILP) is also relevant to this problem, where Liakata et al proposed a methodology, based on ILP, for learning rules which capture what commonly happens (or does not happen) in some domain of interest, using weighted finate state automation (FSA) techniques for the graphical representation. Finally, in the literature is not addressed yet is the problem of finding the type of connectivity between the instances of two semantically related concepts, which in our opinion is crucial.

In summary, our research develops a set of methods for automatic discovery of semantic relations, discovery the type of connectivity between the instances of the related concepts and also infer a set of rules that define semantically the concepts.

Chapter 3

Discovery of Semantic Relations

3.1 Introduction

This chapter presents our proposed approach for discovering semantic relations between concepts from text corpora. Specifically, it presents a statistical methodology that is able to discover directed semantic relation (taxonomic and non-taxonomic) between a set of concepts that have been annotated in a domain specific corpus. It presents the theoretical background, upon which we were based on in order to develop our method, as well as and the requirements that our method has. Moreover, it explains the two different types of concepts, High-Level and Low-Level that we use in our research and also the applicability of our technique in both types of concepts.

3.2 Basic Assumption

Motivated from the constraints mentioned before 1.2.2 and trying to develop an automatic and domain independent method, we have developed a statistical methodology, applied on metadata extracted from the corpus. We adopted a statistical approach in order to face both constraints arisen from text processing as well as to overcome the lack of domain independent techniques for knowledge acquisition from text.

The discovery process of semantic relations is not based on commonly used assumptions that are used in the literature like, that verbs typically indicate semantic relations between concepts or does not exploit syntactic-patterns or clustering methods or any external knowledge-base like WorldNet. Our approach is based on the assumption that concepts which are semantically related, tend to be "near" as context in a plain text. This assumption arises from the principle of coherence on linguistics[34]. Based on this assumption, statistical methods are applied to metadata extracted from the annotated

The 34-year-old, World marathon record holder and two-time Olympic and four-time World 10,000m champion Haile Gebreselassie of Ethiopia today announced that he intends to compete in this 2008 FKB-Games - IAAF World Athletics Tour - in Hengelo, the Netherlands on 24 May in his bid to make Ethiopia's team for the Beijing Olympics in China.

Athlete (name: *"Haile Gebreselassie"*, age: *"34"*, nationality: *"Ethiopia"*, gender: *NotFound*)
SportsCompetition (sport-name: *"10,000m"*, city: *"Hengelo"*, stadium-name: *NotFound*, date: *"24 May"*)

Figure 3.1: Text annotated with instances of two High-Level concepts.

texts, to discover semantic directed relations between concepts as well as to find the type of connectivity on concepts' relations, based on the same assumption too.

3.3 Requirements

As noted before, the method does not require any supervision or any training process. Its only requirement, in order to discover the ontological relational schema of a domain, is the annotation of a corpus on this domain with instances of the ontology's concepts. In other words, its only requirement is a corpus annotated with instances of the concepts that we want to discover their semantic relationships. Figure 3.1, shows an example with a segment of text, annotated with instances of the concepts *Athlete* and *SportsCompetition*.

As you can notice, the two concepts are constituted by a number of attributes, e.g the *Athlete* concept is constituted by the attributes *name*, *age*, *nationality* and *gender*. So, the annotation of a concept's instance includes the process of finding fillers for these attributes, ideally for all. But is not necessary to find fillers for all concept's attributes in the text, in order to annotate an instance, because sometimes this information does not exist explicitly in the text. Probably, some of the attributes' fillers are mentioned implicitly. In general, an instance of a concept is annotated when specific attributes fillers are found that contain "enough", according to the annotator's judgement, semantic information. As shown at Figure 3.1, the athlete's instance does not contain a filler for the attribute gender, nevertheless the annotator has judged that he found "enough" information.

These annotations is the prior knowledge for our proposed method, the rest of the method is automatic and unsupervised. As you will see in the next session 3.4, in our research we categorize the concepts according to the number of the attributes from which they are constituted. Thus, except from the "compound" concepts that are constituted from more than one attributes, our method is also able to discover semantic relations among "simple" concepts that are constituted from only one attribute. The annotation of a corpus with instances of these concepts can be an easy process and automatic, without human annotators that is used for the annotation of a corpus with instances of "compound" concepts.

3.4 High-Level and Low-Level Concepts

In our approach we distinguish the concepts in two types, as **High-Level** and as **Low-Level** concepts depending on complexity of the domain that they represent.

More specifically, we categorize as **Low-Level** concepts, the concepts where their instances are associated with relevant text portion in the document and are directly identifiable. More specifically, we characterize as Low-Level the concepts that are constituted from only one attribute, the attribute *"has-instance"*. Figure 3.2, shows an example of the same, with figure 3.1, segment of text annotated with instances of Low-Level concepts. As you can observe, these concepts are Low-Level because they are constituted from only one attribute ("has-instance") and their instances are associated with relevant text portion in the document, which is directly identifiable.

On the other hand, we categorize as **High-Level** the concepts that are "compound" in such a way that instances of these concepts are related to instances of Low-Level concepts. The High-Level concepts are constituted from more than one attributes and their instances are not directly identifiable in a document. As we previously mentioned in section 3.3, instances of these concepts are sets of fillers, for all or for combinations, of the concept's attributes. Figure 3.1 presents an example of two High-Level concepts, the concepts *Athlete* and *SportCompetition*. Observing the "constraction" of these High-Level concepts you can see that are constituted from more than one attributes, e.g the High-Level concept *Athlete*, is constituted from the attributes *name*, *age*, *nationality* and *gender*. Their instances also, are related to instances of Low-Level concepts.

Our proposed method is both applicable for the discovery of semantic relations between High-Level and also between Low-Level concepts. The application of the proposed method on High-Level concepts is presented in section 3.5, whereas the application on Low-Level concepts in section 3.6.

The 34-year-old, World marathon record holder and two-time Olympic and four-time World 10,000m champion Haile Gebreselassie of Ethiopia today announced that he intends to compete in this 2008 FKB-Games - IAAF World Athletics Tour - in Hengelo, the Netherlands on 24 May in his bid to make Ethiopia's team for the Beijing Olympics in China.

Age (has-instance: *“34”*)
Sport-Name (has-instance: *“marathon”*)
Sport-Name (has-instance: *“10,000m”*)
Name (has-instance: *“Haile Gebreselassie”*)
Nationality (has-instance: *“Ethiopia”*)
Date (has-instance: *“2008”*)
City (has-instance: *“Hengelo”*)
Date (has-instance: *“24 May”*)
Nationality (has-instance: *“Ethiopia's”*)
City (has-instance: *“Beijing”*)

Figure 3.2: Text annotated with instances of Low-Level concepts.

3.5 The proposed methodology for Relation Discovery between High-Level concepts

The proposed methodology for discovering the semantic relation between High-Level concepts involves 3 major steps:

1. Finding the offsets of the annotated instances in the corpus collection.
2. Finding per document the different pairs of concepts that have overlapping instances.
3. Finding the related concepts using the Semantic-Correlation metric.

Detailed explanation for each step is presented in the three sub-sections, respectively.

3.5.1 Finding Instances' Offsets

As noted previously, our approach is based on the assumption that concepts which are semantically related tend to co-occur “near” each other in a plain text, i.e., spatial proximity in text implies semantic similarity. Based on this assumption, we treat each document of the corpus as a sequence of symbols. We consider as symbols all the characters, including spaces and the punctuation marks that exist in the document. In this manner, each document is represented in a one-dimensional Euclidean space, depending on the place in which each symbol is found in the text. We have adopted this transformation in order to represent the documents in a normalized and countable "space". For example, the phrase "*The 34-year-old, World marathon record holder*" is represented with the set $[0, 44]$ because the text is a sequence of 45 symbols. This set that represents a text according to the aforementioned transformation, we call it **offset**. In the same example, the offset of the phrase "*34-year-old, World marathon*" is the set $[4, 30]$, since the phrase starts from the 4^{th} symbol and ends at the 30^{th}.

Based on the aforementioned transformation of the documents, we first find for each document the offsets of the annotated High-Level concepts' instances. As mentioned in the previous section, each instance is formed of the fillers of the concept's attributes found in the text. Consequently, the **offset of an instance** is defined as the range from the first to the last symbol of the instance's fillers. In other words, the offset of an instance is defined as the set of the minimum segment of text which encloses all its fillers. For example, in the document shown at Fig.3.1, whose offset is the set $[0, 342]$ (the text is a sequence of 343 symbols), the offset for the Athlete's instance is the set $[4, 134]$, since it is the minimum part of text which encloses all its fillers or the range of the instance's fillers is from the 4^{th} symbol(filler *"34"* starts from the 4^{th} symbol) to the 134^{th} symbol(filler *"Ethiopia"* ends at the 134^{th} symbol).

3.5.2 Finding Different Pairs of Concepts that have Overlapping Instances

The next step, after finding the offsets of the annotated instances, is to search per document for the different pairs of concepts that have overlapping instances. Specifically, for every document doc_z of the corpus, on which has been annotated instances of different concepts e.g. $C_i, C_j, \ldots, C_n$, where each of these instances has an offset $I_k = [l, r]$. We examine the offset of each instance I_x of C_i, with the instances' offsets of the rest concepts, except C_i's, that have been annotated in the document doc_z, if have overlapping sets. If we find that an instance $I_x \in C_i$ and an instance $I_y \in C_j$ have overlapping sets, then we create for doc_z a pair of concepts $\Big(C_i, C_j\Big)$. This created pair of concepts denotes that the concepts C_i and C_j have in document doc_z overlapping instances. We continue the process for each different annotated concept in a document, for each document of the corpus.

The precise mathematical description of the aforementioned process is presented below:

For the document doc_z*, of the corpus:*

$C_{doc_z} = \{C_1, C_2, \ldots, C_n\} \quad where \quad C_i = \{I_1, I_2, \ldots, I_m\}$

$where \quad I_k = [l, r] \bigcap \mathbb{N} \quad and \quad l < r,$

we compare the instances' offsets:

$\forall(I_x, I_y) \quad where \quad I_x \in C_i, \quad I_y \in C_j$

$and \quad C_i \in C_{doc_z} \quad and \quad C_j \in C_{doc_z} - \{C_i\}$

$$I_x \bigcap I_y \neq \emptyset \quad then\ create\ a\ pair \left(C_i, C_j\right) for\ doc_z \tag{3.5.1}$$

Where:

C_{doc_z}*: the set with the different concepts that have been annotated at least once in the document* doc_z

C_i*: the set with the instances of concept* C_i*, which have been annotated in the document* doc_z

I_k*: the offset of instance* I_k

For each document, a list of concept pairs is created according to (3.5.1). An important note that we must mention here, is that for each document we are interested only in finding the different pairs of related concepts and not the number of occurrences (or overlapping co-occurrence) for each of these pairs. In other words, if we find in doc_x that the instances $I_i \in C_k$ and $I_j \in C_m$ have overlapping offset, then we will create for the doc_x the pair $\left(C_k, C_m\right)$, only if the list with related concepts for this document does not contain this pair. If the list of doc_x already contains it, then we ignore this pair. This is a type of normalization that we do in order to reduce the "noise" of the frequently occurring concepts' instances. According to this note, it is easy to understand that the created list, of concept pairs per document, is symmetric.

Thus, after applying (3.5.1) to the corpus, we create a list per document with the unique different pairs of the related concepts.

3.5.3 The Semantic-Correlation Metric

Then, in order to find the semantic directed relations between concepts, we propose the semantic-correlation metric $S(C_i \to C_j)$ between two concepts C_i and C_j. This metric measures the tendency of concept C_i to be semantically related, either taxonomically or non-taxonomically, with concept C_j, but not the inverse.

The semantic-correlation metric (3.5.2), is defined as the product of the conditional probability $P(C_j|C_i)$ with the sum of the mutual information measure $I(C_i, C_j)$ plus

1. This definition is motivated by our initial assumption that concepts which are semantically related, tend to co-occur "near" each other in a plain text. Therefore, concepts whose instances offsets overlap frequently tend to be semantically related. For the above reason we use in our metric the conditional probability $P(C_j|C_i)$, in order to find for the concept C_i the most probable concept C_j with which has overlapping instances offsets. Furthermore, the mutual information measure [6] is used in order to enhance our metric with the association between the concepts C_i and C_j. If there is a strong association between C_i and C_j, then the conditional probability $P(C_j|C_i) \gg P(C_i) \cdot P(C_j)$, and consequently $I(C_i, C_j) \gg 0$. If there is no interesting association between C_i and C_j, then $P(C_j|C_i) \approx P(C_i) \cdot P(C_j)$, and thus, $I(C_i, C_j) \approx 0$. If C_i and C_j are not associated, then $P(C_j|C_i) \ll P(C_i) \cdot P(C_j)$, forcing $I(C_i, C_j) \ll 0$. Consequently, the high the semantic-correlation score between two concepts is, the more the concepts are related.

We estimate the probabilities by treating each of the different concepts, which have been annotated in the corpus, as a different event and the extracted pairs of related concepts per document (3.5.1) is the set of our observations for the different events. We use maximum likelihood estimation to estimate the probabilities of events(3.5.3), by counting event frequencies in the set of documents.

$$
\begin{aligned}
S(C_i \to C_j) = P(C_j|C_i) \cdot \Big(1 + I(C_i, C_j)\Big) \Leftrightarrow \\
S(C_i \to C_j) = P(C_j|C_i) \cdot \left(1 + log\left(\frac{P(C_j|C_i)}{P(C_i) \cdot P(C_j)}\right)\right)
\end{aligned}
\tag{3.5.2}
$$

$$
P(C_j|C_i) = \frac{P(C_i, C_j)}{P(C_i)} \Rightarrow P(C_j|C_i) = \frac{\frac{\#appearances\; of\; pair\; (C_i, C_j)}{\#all\; pairs}}{\frac{\#appearances\; of\; pair\; (C_i, *)}{\#all\; pairs}} \tag{3.5.3}
$$

Now, in order to find for a concept C_i the concept with which is semantically related (either taxonomically or non-taxonomically), we compute using our proposed metric the semantic-correlation scores between C_i and each of the rest of the concepts. The concept that maximizes this score (3.5.4) is the concept with which the concept C_i is related to.

Find how concepts are related:

$$
C_{corpus} = \{C_1, C_2, \ldots, C_n\}\,, \quad \forall C_i \in C_{corpus},
$$

$$
RELATE \quad C_i \to C_j, \quad \arg\max_{C_j} S\Big(C_i \to C_j\Big), \tag{3.5.4}
$$

$where\ C_j \in C_{corpus} - \{C_i\}$

Applying the aforementioned methodology to the annotated corpus, we manage to find the directed semantic relations between the annotated concepts. The proposed method does not use any lexicon-syntactic patterns and clustering methods, or any external knowledge like WorldNet. We simply apply statistical methods to document metadata that is, to the location of concept instances in text.

3.6 The proposed methodology for Relation Discovery between Low-Level concepts

The previous section showed our proposed methodology for discovering semantic relation among High-Level concepts that have been annotated in a corpus. This section presents the applicability of our approach also in Low-Level concepts. As we mentioned in section 3.4, Low-Level are concepts that are constituted from only one attribute[1] and are directly identifiable in the document. In order to find directed semantic relations among Low-Level concepts we apply the proposed methodology as before, with a variation on the definition of the instance offset of each Low-Level concept. Specifically, we extend the offset of each instance by X symbols to the left and to the right.

According to the definition of the offset instances mentioned in section 3.5.1, the offset of an instance is defined as the range from the first to the last symbol of the instance's fillers and because the Low-Level concepts are constituted from only one attribute, the offset is the set that represents the filler of the unique attribute. We reform that definition for the offset of the Low-Level concepts' instances, by extending the set that represents the attribute's filler by X symbols to the left and to the right. The factor X, is a predefined constant variable.

For example, in figure 3.2, concerning the *Nationality* Low-Level concept, the offset of its instance (*"Ethiopia"*), with a window size X is the set $[(127 - X), (134 + X)]$ (the filler starts from the 127^{th} symbol and ends at the 134^{th}). So, if the window size is $X = 10$ symbols, then this instance offset will be $[117, 144]$. The usage of a window size, is motivated by the fact that instances of Low-Level concepts contain very few words, mainly only one, thus semantically related concepts might be near each other in the text but not overlapping.

The rest of the method for Low-Level concepts is exactly the same, as mentioned before! Due to the fact that the Low-Level concepts are constituted from only one

[1] Low-Level concepts, are concepts that are constituted from only one attribute, the attribute *"has-instance"*

attribute and are directly identifiable in the document, the annotation of a corpus with instances of Low-Level concepts can be an easy and an automatic process. Automatic because nowadays, already exist effective, in terms of precision and recall, systems and methods for terms extraction, e.g, a name entity recognizer. We believe that our proposed approach is very useful for knowledge acquisition from text resources. Especially, for the Low-Level concepts where as we said, using terms extraction systems you are able to annotate automatically the corpora and then applying the proposed methodology to discover the semantic network of the concepts.

3.7 Summary

In this chapter we presented an automatic and domain independent method for discovering semantic relation, both taxonomic and non-taxonomic, between concepts.

Our approach is based on an assumption that concepts which are semantically related, tend to be "near" as context in a plain text. Motivated from this assumption we developed a statistical technique, which applied to documents metadata is able to discover directed semantic relation. The prior knowledge for our methodology is a text corpus annotated with instances of the concepts that we want to discover their semantic relationships. In our research we distinguish the concepts in two types, as High-Level and as Low-Level concepts according to complexity of the domain that they represent. Low-Level, are the concepts where their instances are associated with relevant text portion in the document. On the other hand, High-Level are the concepts that are "compound" in such a way that instances of these concepts are related to instances of Low-Level concepts.

The process for discovering the semantic relation between High-Level concepts involves 3 major steps. The first step involves finding the offsets of the annotated instances in the corpus. Where as offset of an High-Level instances we define the set of the minimum part of text which encloses all its fillers. The next step, is to search per document for the different pairs of concepts that have overlapping instances, in order to create a list per document with the unique different pairs of related concepts. The last step of our proposed methodology, displays the process of finding the related concepts using the Semantic-Correlation metric. The Semantic-Correlation metric between two concepts A and B, is a score metric based on statistics that measures the tendency of concept A to be semantically related, either taxonomically or non-taxonomically, with concept B, but not the inverse. In order to discover for a concept A the concept with which is semantically related, we compute the semantic-correlation scores between A

and each of the rest of the concepts. The concept that maximizes this score, is the concept with which the concept A is related to.

In order to find directed semantic relations among Low-Level concepts we apply the proposed methodology, with a variation on the definition of the instance offset of each Low-Level concept. The variation on the definition of the Low-Level instance's offsets is by using a window, extending the offset of each instance by X symbols to the left and to the right.

Chapter 4

Inferred Knowledge about the Domain of the Extracted Ontology

4.1 Introduction

The last component, in the process of ontology learning from textual resources, is the task of discovering a set of inference rules and/or constraints about the domain of the extracted ontology. This knowledge is important and also necessary because "describes", in a machine readable way, the implicit knowledge of domain. This chapter presents two methods for finding two types of inferred knowledge about the extacted ontology. We propose a methodology for finding the type of connectivity among the instances of two related concepts. Furthermore, we present an algorithm for learning a minimum set of rules for the High-Level concepts, where these rules define semantically the concepts in the level of the semantic info that their instances must contain.

.

4.2 Discovering the Type of Connectivity of Related Concepts

Apart from the discovery of the semantic relations between ontology concepts, another major task in the field of ontology learning which, as far as we know, is not addressed yet is the problem of finding the type of connectivity between the instances of two semantic related concepts. We believe that this knowledge is very important, because offers "more information" for more effective computational logical inference on the ontology.

We present an automatic statistical method, which is able to discover the type of connectivity among the instances of the related concepts that we have discovered with the methodology presented in the previous chapter. The types of connectivity among two related concepts, that the proposed methodology is able to specify, are $1 : N$ (one-to-many), $N : 1$ (many-to-one) and $M : N$ (many-to-many). We find the type of connectivity between two concepts, based on the initial assumption that concepts, whose instances' offsets overlap, tend to be related. We extend this syllogism by investigating the way that the instances' offsets are overlapped, in order to discover the type of their relationship. Hence, we specify as type of connectivity between the instances of two related concepts the type which occurs more often in the corpus.

The proposed methodology, for discovering the type of connectivity between instances of related concepts $C_A \rightarrow C_B$, consists of the following steps:

1. We find all the documents $DOC_{C_A\_and\_C_B} = \{doc_1, doc_2, \ldots, doc_n, \}$ in the corpus that contains instances of the concepts C_A and C_B.

2. For each document doc_z that contains instances of concepts C_A and C_B, we create a list per document doc_z with the overlapping instances of the concepts C_A and C_B. Specifically:

   $\forall doc_z \in DOC_{C_A\_and\_C_B}, \quad where \quad C_{doc_z} = \{\ldots, C_A, \ldots, C_B, \ldots\},$
   $where \quad C_A = \{I_{A_1}, I_{A_2}, \ldots, I_{A_n}\} \quad and \quad C_B = \{I_{B_1}, I_{B_2}, \ldots, I_{B_M}\},$
   $where \quad I_{A_x} \ or \ I_{B_y} = [l, r] \bigcap \mathbb{N} \quad and \quad l < r,$
   $we\ compare\ the\ instances'\ offsets\ \forall (I_{A_i}, I_{B_j}) :$

   $$If\Big(I_{A_i} \bigcap I_{B_j} \neq \emptyset\Big) \quad then\ create\ a\ pair\ \Big(I_{A_i}, I_{B_j}\Big)\ for\ doc_z \qquad (4.2.1)$$

   Where:
   C_{doc_z}*: the set with the different concepts that have been annotated in the document* doc_z*, which must also contains the concepts* C_A *and* C_B
   C_A*,* C_B*: the set with the instances of concepts* C_A*,* C_B*, which have been annotated in the document* doc_z
   I_{A_x}*,* I_{B_y}*: the offset of instance* I_{A_x}*,* I_{B_y}

   Note in this step, in contrary with the methodology for discovering the relations, we are interested in finding all overlapping instances per document and not only the different pairs.

3. After creating the list with overlapping instances of the concepts C_A and C_B for doc_z, we find the type of connectivity, for doc_z, between the instances of these concepts as follows:

$$If \quad \left.\begin{matrix} I_{A_i}, I_{B_j} \\ I_{A_i}, I_{B_m} \\ I_{A_i}, I_{B_n} \\ \dots \end{matrix}\right\} Then\ the\ type\ of\ connectivity\ for\ doc_z\ is\ \big(1:N\big)$$

$$Else-If \quad \left.\begin{matrix} I_{A_i}, I_{B_j} \\ I_{A_k}, I_{B_j} \\ I_{A_m}, I_{B_j} \\ \dots \end{matrix}\right\} Then\ the\ type\ of\ connectivity\ for\ doc_z\ is\ \big(N:1\big)$$

$$Else \quad \left.\begin{matrix} I_{A_i}, I_{B_j} \\ I_{A_j}, I_{B_k} \\ I_{A_m}, I_{B_j} \\ \dots \end{matrix}\right\} Then\ the\ type\ of\ connectivity\ for\ doc_z\ is\ \big(M:N\big)$$

We continue the aforementioned steps 2 and 3 for all documents of $DOC_{C_A\_and\_C_B}$. Thus, after finishing these steps we will have found for each document of $DOC_{C_A\_and\_C_B}$, the type of connective between the instances of the concepts C_A and C_B.

4. We specify as type of connectivity, for the related instances of concepts C_A and C_B, the type of connectivity that occurs more often in the corpus. In other words, we specify as type of connectivity, the more frequent type of connectivity among the documents that contain overlapped instances of concepts C_A and C_B.

4.3 Rules that Describe Semantically Ontology's Concepts

Apart from discovering the type of connectivity between the instances of related concepts, we also presents another type of inferred knowledge about the extracted ontology. We propose an algorithm $O(2^n)$ that is able to learn for a High-Level concept a set of rules, which describe the minimum semantic info that an instance must "contain".

As we mentioned in section 3.3, a High-Level concept is constituted from more than one attributes and the annotation of an instance includes the process of finding fillers for these attributes, ideally all. But is not necessary to find fillers for all concept's attributes in the text, in order to annotate an instance, because sometimes this information does not exist explicitly in the text. In general, an instance of a concept is annotated when specific attributes fillers are found that "contain" enough, according to the annotator's judgement, semantic information. Our proposed algorithm tries to sketch annotator's judgement according to the attribute's fillers that an instance must contain. In other words we present an algorithm, which based on set theory finds rules for a High-Level concept, where these rules define the "minimum semantic" information in the level of attributes' fillers, that an instance must contains.

The proposed algorithm for a High-Level concept C_A, is outlined below:

1. for the concept C_A, a list L is created containing all the different combinations of the concept's attributes with which annotated instances of this concept appear in the corpus.

2. the elements of L are examined for common attributes. If such attributes exist, a rule is created, which denotes that each instance of C_A must contain fillers for these attributes.

3. if we found common attributes, then these are removed from the elements of L.

4. a set S is created containing the remaining different attributes from all the elements of L. The set O is then created having as elements the power-set[1] of S, except from the empty set, $O = P(S) - \{\emptyset\}$. The set O now, contains as elements all possible subsets of S.

5. for each element X of set O ($\forall X \in O$), we examine if X is found in some elements of L.

   - if X is found in some elements of L, we examine the rest elements of L to find a common set of attributes Y (Y must not contain attributes that are in X);

     - if Y is found, then we create a candidate complementary rule which denotes that an instance of C_A may contain fillers for the X or for the Y set of attributes
     – if Y does not exist, then we check L to find an element that is constituted precisely from X. If L contains such an element, then we create a rule, which denotes that each instance of C_A is able to be defined, if it contains

[1] The power-set $P(S)$ in this step, is a factor that causes exponential complexity $O(2^n)$ in our algorithm. If S is a finite set with $|S| = n$ elements, then the power-set of S contains $|P(S)| = 2^n$ elements.

fillers for the X set of attributes. We create this rule, when no other rule, which is subset of X, already exists!

6. finally, we collect all the candidate complementary rules that we have found and delete those rules that are subsets of other complementary rules.

To demonstrate the proposed algorithm, we present below an example of the rules that the algorithm discovers for the concept C_A.

If the concept C_A is constituted from the following attributes:
$(A, B, C, D, E, F, G, H, I, J, K)$

and the list L with the different combinations of the instances' annotations for C_A in the corpus are:
(A, B, C, D, E, F, H, K)
(A, B, D, H, K)
(A, B, C, E)
(A, D, F, G, H)
(A, B, E, G, I, J)
(A, D, K)

The discovered rules for the minimum semantic information that the concept C_A must contain are:

$$\left(A \quad \text{AND} \quad \left(D \quad \text{OR} \quad (B \quad \text{AND} \quad E) \right) \right)$$

Evaluating the acquired rules to the list L, with the different combinations of the instances' annotations, we observe that these rules express the least common subset of the attributes, which occurred in every instance of C_A. Because each instance contains at least fillers for the attributes, A and either for D or for B and E.

4.4 Summary

In this chapter we proposed two automatic methods for finding inferred constrains and rules about the domain of an extracted ontology. This knowledge is important because "describes", in a machine readable way, the implicit knowledge of domain, making easier the logical inference for a computational system.

We presented an automatic statistical method, which is able to discover the type of connectivity among the instances of the related concepts. The types of connectivity that the proposed methodology is able to specify, are $1 : N$ (one-to-many), $N : 1$ (many-to-one) and $M : N$ (many-to-many). Motivated from our initial assumption that concepts, whose instance offsets overlap, tend to be related. We extend this syllogism by investigating the way that the instances' offsets are overlapped, in order to discover the type of their relationship. Based on this, we specify as type of connectivity between the instances of two related concepts the type which occurs more often in the corpus.

Moreover, we presented another method for finding another type of inferred knowledge. We proposed an algorithm, based on set theory, that is able to learn for a High-Level concept a set of rules, which describe the minimum semantic info that an instance must contain.

Chapter 5

Evaluation

5.1 Introduction

In the previous two chapters, we presented methods for discovering semantic relations and inference rules for a set of concepts, from textual corpora annotated with instances of these concepts. Applying these methods to a corpus annotated with instances of either High or Low Level concepts, we are able to extract an ontology that will be formed from the annotated concepts. In this chapter we present the experimental results of our methods, applied on two corpora of different domains and the extracted ontologies were evaluated with respect to the corresponding manually created ontologies.

5.2 Experimental Corpora

The experimental corpora are from the athletics and biomedical domain and contain annotations for both High and Low Level concepts. The corpus on athletics domain was obtained from the EC-funded project BOEMIE[1]. The second corpus on biomedical domain is from abstracts of Pubmed[2] on allergens.

We did not use corpora, like GENIA[3], due to type of annotation our approach requires as prior knowledge. Our method uses annotated concepts' instances which are formed of the fillers of concepts attributes. Consequently, the annotated concepts need to have more than one attribute, which was not the case in the publicly available corpora we examined.

[1] http://www.boemie.org

[2] http://www.ncbi.nlm.nih.gov/pubmed/

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/ genia/topics/Corpus

5.2.1 Boemie Corpus

The first corpus is from the athletics domain and consists of 2087 web pages, with content, collected mainly from the IAAF[4] web site. This corpus contains instances' annotations for 20 different High-Level concepts, where these concepts are formed from 29 different attributes.

It contains 36,240 instances annotations for 20 High-Level concepts and also 56,494 instance annotations for 13 Low-Level concepts. This corpus has already been used in [10], [31]. The corpus documents contain athletic articles for 10 different sports competitions. A part of the manually created ontology, containing the annotated High-Level concepts, developed in the context of the same project by human experts, is presented in figure 5.1(a).

The 20 High-Level concepts with their attributes that have been annotated in the corpus are:

Athlete *(has-Name, has-Age, has-Gender, has-Nationality)*

MaleAthlete *(has-Name, has-Age, has-MaleGender, has-Nationality)*

FemaleAthlete *(has-Name, has-Age, has-FemaleGender, has-Nationality)*

SportsRound *(has-RoundName, Starting-date, Finishing-date)*

SportsEvent *(has-EventName, In-City, On-Country, Starting-date, Finishing-date)*

SportsTrial *(has-Performance, has-Ranking)*

SportCompetition *(has-SportName, InCity, Starting-date, Finishing-date, In-StadiumName)*

JumpingCompetition *(has-JumpingName, InCity, Starting-date, Finishing-date, In-StadiumName)*

ThrowingCompetition *(has-ThrowingName, InCity, Starting-date, Finishing-date, In-StadiumName)*

RunningCompetition *(has-RunningtName, InCity, Starting-date, Finishing-date, In-StadiumName)*

TripleJumpCompetition *(has-TripleJumpName, InCity, Starting-date, Finishing-date, In-StadiumName)*

PoleVaultCompetition *(has-PoleVaultName, InCity, Starting-date, Finishing-date, In-StadiumName)*

[4] http://www.iaaf.org

HighJumpCompetition *(has-HighJumpName, InCity, Starting-date, Finishing-date, In-StadiumName)*

LongJumpCompetition *(has-LongJumpName, InCity, Starting-date, Finishing-date, In-StadiumName)*

HammerThrowCompetition *(has-HammerThrowName, InCity, Starting-date, Finishing-date, In-StadiumName)*

JavelingThrownCompetition *(has-JavelingThrownName, InCity, Starting-date, Finishing-date, In-StadiumName)*

HurdlingCompetition *(has-HurdlingName, InCity, Starting-date, Finishing-date, In-StadiumName)*

Running100mCompetition *(has-Running100mName, InCity, Starting-date, Finishing-date, In-StadiumName)*

MarathonCompetition *(has-MarathonName, InCity, Starting-date, Finishing-date, In-StadiumName)*

RaceWalkingCompetition *(has-RaceWalkingName, InCity, Starting-date, Finishing-date, In-StadiumName)*

The 13 Low-Level concepts, which are the fillers for the High-Level concepts' attributes that have been annotated in the corpus are:

Name, ***Gender***, ***Age***, ***Nationality***, ***Performance***, ***Ranking***, ***Sport-Name***, ***Round-Name***, ***Stadium-Name***, ***Event-Name***, ***Date***, ***City***, ***Country***

5.2.2 Allergen Corpus

The second corpus is from the biomedical domain and consists of 183 abstracts of Pubmed[5] on allergens. It contains instances' annotations for 6 different High-Level concepts that are formed from 10 different attributes.

This corpus, contains 1,230 instances annotations for 6 different High-Level concepts and also 2,646 instance annotations for 10 Low-Level concepts. The allergen corpus has also been used in [46]. The manually created ontology from human experts is depicted in figure 5.2(a).

The 6 High-Level concepts with their attributes that have been annotated in the corpus are:

[5] http://www.ncbi.nlm.nih.gov/pubmed/

Allergens *(has-AllergenNameCommon, has-AllergenNameScientific, has-IsoelectricPoint, has-MolecularWeight, is-MajorORMinor)*

Protein *(has-ProteinFamily, has-ProteinName)*

Allergie *(has-AllergenGroup)*

Allergen Sources *(has-SourceCommonName, has-SourceScientificName)*

Named Allergens *(has-AllergenNameCommon, has-AllergenNameScientific, has-IsoelectricPoint, has-MolecularWeight, is-MajorORminor)*

Descriptive Allergens *(has-AllergenNameCommon, has-IsoelectricPoint, has-MolecularWeight, has-MajorORminor)*

The 10 Low-Level concepts, which are the fillers for the High-Level concepts' attributes that have been annotated in the corpus are:

> ***Allergen Name Common***, ***Allergen Name Scientific***, ***Isoelectric Point***, ***Molecular Weight***, ***Major or Minor***, ***Protein Family***, ***Protein Name***, ***Allergen Group***, ***Source Common Name***, ***Source Scientific Name***

5.3 Experimental Results on Boemie Corpus

This section shows the experimental results of our proposed methods for ontology extraction on Boemie corpus. We have applied our methods on Boemie corpus for both the High and Low level concepts and we demonstrate the extracted knowledge in contrast with the corresponding manually created.

5.3.1 Experimental Results for High-Level Concepts

Applying our method on Boemie corpus for the High-Level concepts, we construct the ontology that is presented in figure 5.1(b). The processing time for this experiment was less than 4 minutes (with Intel Centrino Duo, 1.83GHz and 1G memory).

Discovered Relations

Comparing the two ontological relational schemata you can notice that the manually created in contrast with the automatically created, are very close. Specifically, comparing the relations between the High-Level concepts of two ontologies, our method has

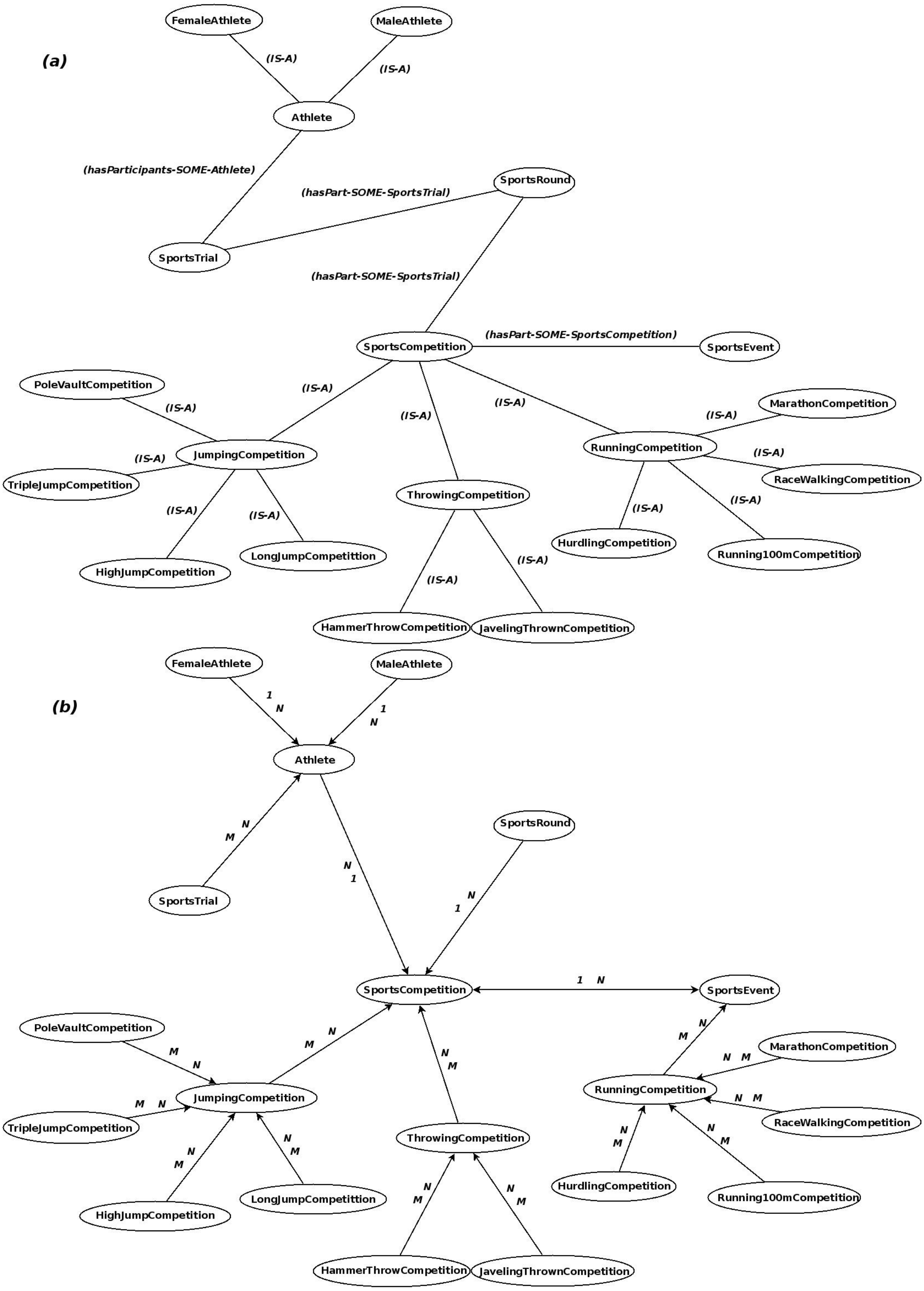

Figure 5.1: (a) The manually created ontology for the domain of athletics. (b) The automatically extracted ontology.

missed only two of the nineteen or 90% expressed in the level of precision measurement. Moreover, observing the manually created ontology, it can be noticed that its relational schema contains both taxonomical (*IS-A*) and non-taxonomical (*hasPart-some-...*) relations. Our method has found relationships of both types and the most of them are correctly.

Our method has missed only two relations, compared with the manually created one. It relates the concept *"Athlete"* with the concept *"SportsCompetition"*, instead of relating the concept *"SportsTrial"* with the concept *"SportsRound"*, which nevertheless is not semantically incorrect. Because, obviously an athlete is related with a sport competition. The other missed relation is among the concepts *"RunningCompetition"* and *"SportsEvents"*, instead of relating it, with the concept *"SportsCompetitions"*. This happened due to the fact that the Marathons' names in most times, in the documents, are mentioned with the city in which they took place (e.g. London-Marathon, Berlin-Marathon,...). That has as effect, the instances' offsets of MarathonCompetition to be overlapped with the instances' offsets of *SportsEvent*, because the *"SportsEvent"* concept has the attribute city. We have evaluated our method without the MarathonCompetition's instances in the corpus and the *"RunningCompetition"* concept was related correctly, with the *"SportsCompetition"* concept. Appendix A.1, contains the semantic-correlation scores between *"RunningCompetition"* concept and and each of the rest of the concepts of the Boemie corpus, with and without the MarathonCompetition's instances. As you can notice from the list of scores, even when our method relates incorrectly the concept *"RunningCompetition"* with *"SportsEvent"* instead of relating it with *"SportsCompetition"*, then the scores are almost the same, their difference is $0.006 \approx (-1,5\%)$.

Discovered Inference Knowledge

Applying our proposed methods we also found the type of connectivity for the discovered relations and a set of rules for each concept with the minimum semantic info that its' instances must contain.

Figure 5.1(b), depicts also the type of connectivity found between the instances of the related concepts. The discovered types seems very reasonable, especially for the non-taxonomic relations. For example, the method specified the relation between the concepts *"Athlete"* and *"SportsCompetition"* as of type $(1 : N)$, which is reasonable, because many athletes participate to one sport. Also the relation between *"SportsRound"* and *"SportsCompetition"* as of type $(N : 1)$, which is also reasonable, since one sport has many rounds(final, semi-final), etc. Observing the type of connectivity for the taxonomic relations, you will notice that the discovered type is $(M : N)$. This happens because the percentage of the overlapped offsets' sets between the instances of these concepts, is high. This phenomenon is reasonable to be occurred, because when we have two concepts that are related with a taxonomic relation, the instances' offset for one of these is usually a subset of the other. Consequently, the instances of these concepts are overlapped frequently that's why the type of connectivity is $(M : N)$. In conclusion, from the experimental results we comprehend that we can use the percentage of the overlapped offsets' sets between the instances of related concepts, in order to label the semantic relation as hypernym or hyponym.

Finally, for this experiment, our proposed algorithm has learned for each concept a minimum set of rules, for its attributes, which define it semantically. Due to the fact that the High-Level concepts for this domain are formed only from a few attributes, e.g. 4 attributes. It is difficult to find complicated rules. We mention the most complicated rules that we found. According to this, the minimum information that must be found in a document, in order to define a new instance of concept *"Athlete"* is the fillers for the attributes $\Big($*has-Name* AND $\big($*has-Gender* OR *has-Age* OR *has-Nationality*$\big)\Big)$. While, for the concepts *"Female"* and *"MaleAthlete"* these rules are $\Big($*has-Name* AND *has-FemaleGender/has-MaleGender* AND $\big($*has-Age* OR *has-Nationality*$\big)\Big)$ that are reasonable. Because from the first rule we found that an *"Athlete's"* instance must contains at least the name of the athlete and either his/her age or gender or nationality. But for a *"Female"* or *"MaleAthlete"* instance must at least contains both the name and his/her gender and either his/her age or nationality, because the important attribute in order to categorize an "Athlete's" instance as male or female is his/her gender.

5.3.2 Experimental Results for Low-Level Concepts

Applying our method as presented in section 3.6 on Boemie corpus for the 13 Low-Level concepts for a window size X of 50 symbols, we found the ontological relational schema that is presented in figure 5.2. The processing time for this experiment was less than 2 minutes (with Intel Centrino Duo, 1.83GHz and 1G memory).

Discovered Relations

Observing the discovered relations of the Low-Level concepts, it can be noticed that they seem very reasonable. Unfortunately we do not have a gold standard, from human experts, to compare our results, but it easy to evaluate them because of the domain that they represent. The system relates the concept "performance" with the concept "ranking", that is reasonable. It also relates "round-name" with "sport-name", which is reasonable becuase a round(final, semi-final) is related with a sport. The concept "name" is related with "nationality" and "age" and "gender", the name of person (an

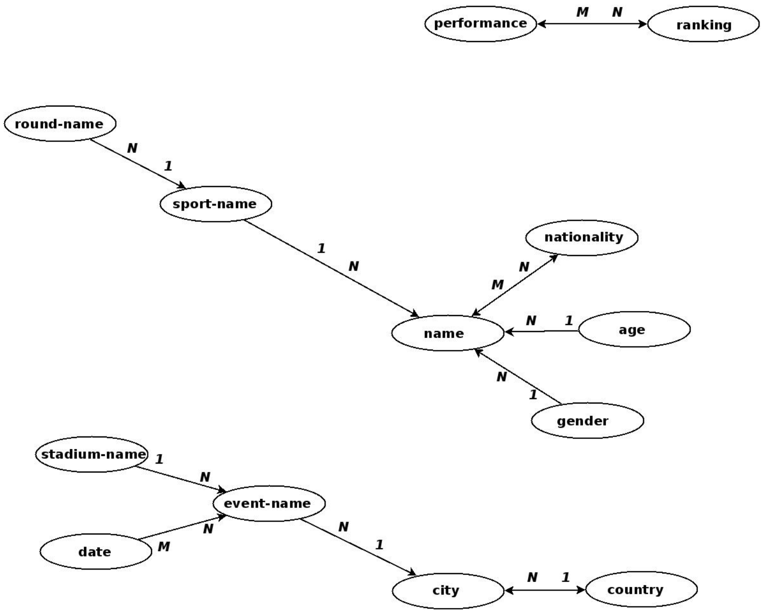

Figure 5.2: The extracted semantic relations among Low-Level concepts, using a window size of 50-symbols.

athlete in our domain) is related with an age, a gender and a nationality. Moreover, it relates the "sport-name" with the "name", which is also obvious, a sport is related with athletes. It relates the concepts "stadium-name" and "date" with the "event-name" concept and an "event-name" with "city" which is also related with "country". We believe that the discovered relational schema for these concepts is realistic and reasonable.

We would like to mention at this point another remarkable observation from the experimental results. Our method relates the concept "round-name" with the concept "sport-name", as we said is reasonable because a final or semi-final round is related with a sport. But as you know a final or semi-final round is also related with the gender of the athlete, women-final or men-semi-final. Due to the way that are our method relates the concepts, each concept is related with the concept that maximize its semantic-correlation score, it cannot relates a concept with more than one concept. Watching the list with the scores A.2 for the "round-name" and the rest of the concepts we observe that the maximum score is $S(round\_name \rightarrow sport\_name) = 0.610$ and the second maximum is $S(round\_name \rightarrow gender) = 0.605$, between "round-name" and "gender". The first from the second differ for $0.005 \approx -0.18\%$. From these observations we comprehend that we can extend the way that our method relates the concepts, not only with the concept that maximize its score but also with the concepts that their scores are very close to maximum score, e.g $-1.5\% or -2\%$ of the maximum score. With this variation we are able to relate a concept with more than one concepts.

The window size (*WS*) for this experiment was 50-symbols. The same results are also discovered for window size 100-symbols. For window size larger than 100-symbols, we observed that all the Low-Level concepts tend to be related with the concept *name*. This is expected since the concept *name* is the more frequently occurring one. In general, the usage of a large *WS* leads to over-generation of semantic relationships as an increasing number of concept instances are now overlapping. From experimentation with the *WS* for different corpora and different Low-Level concepts, we conclude that the best *WS* is related with the density of the annotated concept instances in the text. The rule of thumb is: the higher the density the lower the *WS* should be and vice versa.

Finally, it is also remarkable the fact that the method also "clusters" the Low-Level concepts. As depicted in Fig.5.2, our proposed algorithm has discovered three clusters of related Low-Level concepts. Each of these clusters can be considered as a High-Level concept which consists of Low-Level concepts.

Discovered Inference Knowledge

Applying our proposed methods we also found the type of connectivity for the discovered relations between the Low-Level concepts, but we did not also applied our methodology for rule extraction, because the Low-Level cocnepts are formed from only one attribute (*"has-instance"*).

Figure 5.2, depicts also the type of connectivity found between the instances of the related concepts. The discovered types seems very reasonable, too. Many "round-names" are related with one "sport-name" that is reasonable, because a final, a semi-final round are related with a sport. Also reasonable, many "name" are related with one "sport-name", as well as a "stadium-name" is related with many "event-names" and many "dates" with many "event-names". Finally, many "event-names" are related with one "city" (in a city took place many different sport-events) and many "cities" are related with a country, which is obvious.

Moreover, applied our method on the 10%(200 documents randomly selected) of the corpus with the same window size, we also found the same ontological relational schema as before, both in terms of relations and also on types of connectivities. Which denotes that our approach does not need a "huge" corpus in order to converge to the final ontological relational schema.

In conclusion, we believe that the discovered type of connectivity as well as the discovered relation, for Low-Level concepts, are reasonable and very promising.

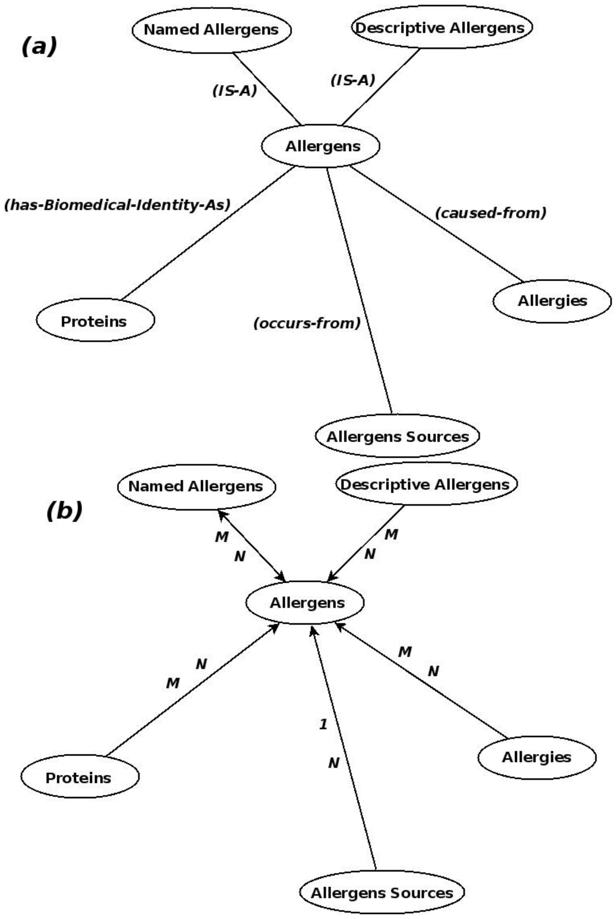

Figure 5.3: (a) The manually created ontology for the domain of allergens. (b) The automatically extracted ontology.

5.4 Experimental Results on Allengen Corpus

This section shows the experimental results of our proposed methods for ontology extraction on Allergen corpus. We have applied our methods on Allergen corpus for both the High and Low level concepts and we demonstrate the extracted knowledge in contrast with the corresponding manually created.

5.4.1 Experimental Results for High-Level Concepts

Applying our method on biomedical corpus for the High-Level concepts, we construct the ontology that is presented in figure 5.3(b). The processing time for this experiment was less than a minute (with Intel Centrino Duo, 1.83GHz and 1G memory).

Observing the automatically constructed ontology, we comprehend that it has exactly the same relational schema among the related concepts with the manual constructed one from human experts. The manually created allergen ontology contains both taxonomic and non-taxonomic relations which our method has discovered correctly. Due to the domain of the ontology, must be an expert on biology or medicine in order to further evaluate the extracted ontology.

Because of the lack of information in work of [46], from which we got the Allergen corpus, we cannot evaluate intensively the inferred knowledge in the level of type of connectivity and also from the process of rules acquisitions. Figure 5.2(b) depicts also the types of connectivity, which have been specified automatically, between the related concepts. We observe that the discovered types are mainly of type $(M : N)$, which happens because the documents of the corpus are small. As we mentioned above, the Allergen corpus consists of abstracts of Pubmed. Consequently, the annotated instances are overlapped each other because of the small region of the document. Observing the only type of connectivity that is not of type $(M : N)$, for the related concepts "Allergens" and "Allergens Sources" that are related with type $(1 : N)$. This type seems reasonable, because an allergen mainly occurs from many and different sources.

The extracted rules, which define for each concept, the minimum information that must exist in order to define a new instance of this concept are:

For the concept of ***Allergens***: ((has-AllergenNameCommon AND (has-IsoelectricPoint OR has-MolecularWeight) AND (has-AllergenNameScientific OR is-MajorORMinor))

For the concept of ***Proteins***: (has-ProteinFamily OR has-ProteinName)

For the concept of ***Allergies***: (has-AllergenGroup)

For the concept of ***Allergen Sources***: (has-SourceCommonName OR has-SourceScientificName)

For the concept of ***Named Allergens***: ((Allergen Name common AND (has-IsoelectricPoint OR has-MolecularWeight) AND (has-AllergenNameScientific OR is-MajorORMinor))

For the concept of ***Descriptive Allergens***: (has-AllergenNameCommon

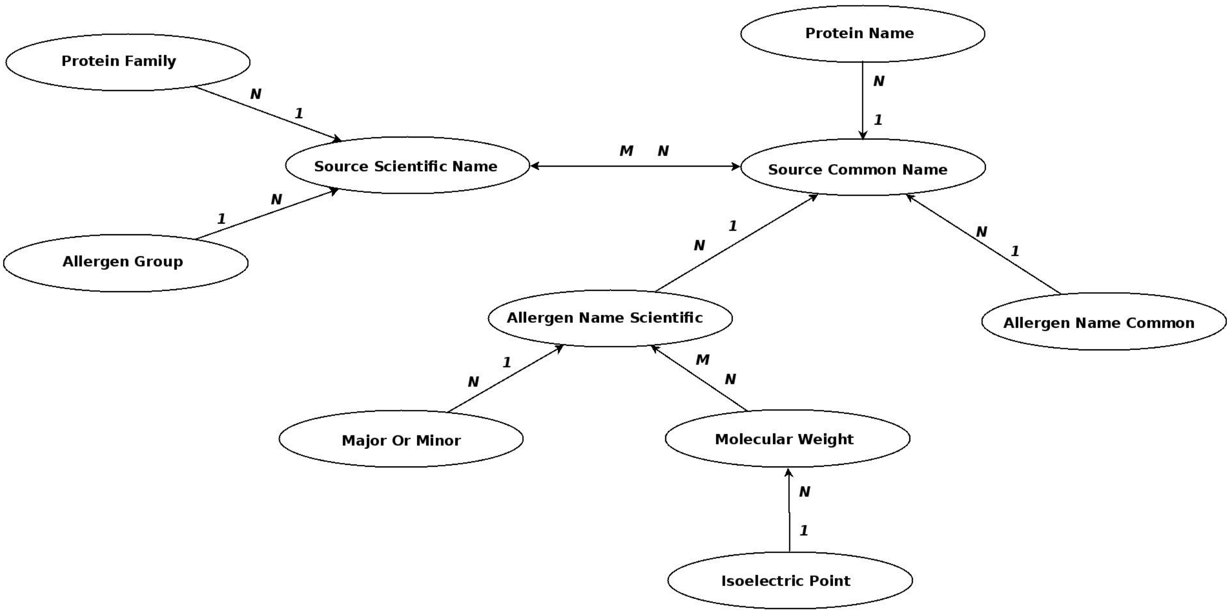

Figure 5.4: The extracted semantic relations for the domain of allergens between the Low-Level concepts, using a window size of 5-symbols.

$$\text{AND }\big(\text{has-IsoelectricPoint OR has-MolecularWeight}\big)\text{ AND is-MajorORMinor}\Big)$$

5.4.2 Experimental Results for Low-Level Concepts

Applying our method as presented in section 3.6 on Allergen corpus for the 10 Low-Level concepts for a window size X of 5 symbols, we found the ontological relational schema that is presented in figure 5.4. The processing time for this experiment was less than a minute (with Intel Centrino Duo, 1.83GHz and 1G memory).

As mentioned before, due to the domain of the ontology and also the lack of a gold standard ontological schema is difficult to evaluate our results. We showed our results in a biologist, who discussed them as reasonable. More specifically, she mentioned that the molecular weight is related with an allergen and the isoelectric point is related with the molecular weight. She also mentioned that the proteins are also related with the sources of allergens.Figure 5.4, depicts also the types of connectivity, which have been specified automatically, between the related concepts.

Finally, one should bear in mind that the evaluation of ontologies when these ontologies are produced by an automated learning procedure is an open field of research. The research community has not established a standard methodology for automating ontology evaluation [48]. Especially when the evaluation is done against a gold standard ontology, it seems that we cannot judge objectively the result, since the gold standard was created by humans probably in a subjective or a biased manner.

5.5 Summary

In this chapter we presented the experimental results of our proposed methods, for automatic extraction of domain specific ontologies from text resources that are annotated with concepts' instances. We demonstrated the experimental results of our methods, applied on two corpora of different domains and the extracted ontologies were evaluated with respect to the corresponding manually created ontologies. The experimental corpora are from the athletics and biomedical domain and contain annotations for both High and Low Level concepts. The corpus on athletics domain was obtained from the EC-funded project BOEMIE, while the second on biomedical domain, is from abstracts of Pubmed on allergens.

The Boemie corpus contains instances' annotations for 20 different High-Level and 13 Low-Level concepts. We have applied our methods on Boemie corpus for both High and Low level concepts and we demonstrated the extracted knowledge in contrast with the corresponding manually created. For the High-Level concepts, the extracted ontological relational schema with the manually created, are very close. Specifically, comparing the relations between the High-Level concepts of two ontologies, our method has missed only two of the nineteen. Applying our proposed methods we found the type of connectivity for the instances of the discovered relations and also a set of rules for each concept, with the minimum semantic info that its instances must contain. The discovered types, as well as the rules seems very reasonable.

Applying our method on Boemie corpus for the 13 Low-Level concepts for a window size X of 50 symbols, we found the ontological relational schema for the Low-Level concepts. Observing the discovered relations of the Low-Level concepts, it can be noticed that they seem very reasonable. From experimentation with the WS for different corpora and different Low-Level concepts, we conclude that the best WS is related with the density of the annotated concept instances in the text. The rule of thumb is: the higher the density the lower the WS should be and vice versa. Finally, it is also remarkable the fact that the method also "clusters" the Low-Level concepts, in a way that each of these clusters can be considered as a High-Level concept which consists of Low-Level concepts.

The Allergen corpus, contains instances annotations for 6 different High-Level concepts and 10 Low-Level concepts. We have applied our methods on Allergen corpus for both High and Low Level concepts and we demonstrated the extracted knowledge in contrast with the corresponding manually created. For the High-Level cocnepts, the automatically constructed ontology has exactly the same relational schema, among the related concepts, with the manual constructed one from human experts. The manually created allergen ontology contains both taxonomic and non-taxonomic relations which our method has discovered correctly. Applying our method for the 10 Low-Level concepts, with a window size X of 5 symbols, we found the ontological relational schema for the Low-Level concepts. Due to the domain of the ontology and also the lack of a gold standard ontological schema, it was difficult to evaluate our results. We showed our results in a biologist, who discussed them as reasonable.

In conclusion, we believe according to experimental results that our proposed novel methods, for knowledge acquisition from text resources, are very promising.

Chapter 6

Conclusion and Future Work

6.1 Introduction

In this chapter we summarize the research contribution of this thesis and we provide useful directions for future work. In this master thesis we proposed a novel methodology for automatic extration of domain specific ontologies from text resources. Our proposed methods using as prior knowledge a corpus annotated with instances of concepts, are able to automatically discover the ontological relational schema between the annotated concept and also to infer knowledge about the domain of the discovered ontology.

6.2 Conclusion

We presented an automatic and domain independent methodology that is able to discover directed semantic relations taxonomic (hierarchical) and non-taxonomic (non-hierarchical) between either "compound" (High-Level) or "simple" (Low-Level) concepts. The discovery process of semantic relations is not based on commonly used assumptions that are used in the literature, that verbs typically indicate semantic relations between concepts or does not exploit syntactic-patterns or clustering methods or any external knowledge-base like WorldNet. Our approach is based on the assumption that concepts which are semantically related, tend to be "near" as context in a plain text. This assumption arises from the principle of coherence on linguistics. Based on this assumption, we apply statistical methods to metadata extracted from the annotated texts, in order to discover directed semantic relations between concepts. Where the metadata are, the overlapping instances of the different concepts in the annotated texts. Moreover, we propose a methodology, based on the same assumption, in order to find the type of connectivity on concepts relations. Finally, we have expounded an algorithm based on set theory that is able to discover a minimum set of rules which define semantically the least info that must exist for the representation of a High-Level concept's instance.

As far as we know, this is the first time in the literature that a domain and language independent methodology for discovering semantic relations between a set of "compound" (High-Level) concepts is proposed. Our method is both domain and language independent, because we do not use any training process or use any syntactic-pattern or grammatical-rule. It is a statistical method that is based on a common, to all languages, principle [34]. Furthermore, with a variation on the definition of the instances' offsets, we also showed the applicability of our approach to "simple"(Low-Level) concepts. Our methodology is automatic, the only step that needs supervision is the step of annotating the corpus with the concepts' instances. As we will present in the next section, an aspect for future work is to further automatize our method by annotating the corpora automatically. Moreover, as mentioned in the previous chapter 5.3.2, our approach does not need a "huge" annotated corpus in order to converge to the final ontological relational schema.

Another aspect that this research demonstrates, which as far as we know is not addressed yet, are methods for extracting inferred knowledge about the domain of the discovered ontology. Our approach except of discovering the ontological relational schema is able to find the type of connectivity between the instances of the related concepts and also to find a minimum set of rules that define semantically the concepts, in the level of the semantic info that their instances must contain. We believe that this knowledge is both important and necessary, because "describes" in a machine readable way the implicit knowledge of domain. Which are useful information for a more efficient and accurate logical inference from computational systems.

The proposed method was applied to corpora from two different domains, athletics and biomedical, and was evaluated against the existing manually created ontologies for these domains. It was applied for both types of concepts (High-Level and Low-Level) and the results proved to be very promising in both domains and for both types of concepts.

In conclusion, it seems from the experimental results that the initial assumption for discovering semantic relation based on the overlapping instances is correct, but it needs further research investigation.

6.3 Future Work

This section demonstrates our future plans on this research field. As mentioned before, this a novel approach that seems to be very promising but it needs further investigation. Future work focus on two directions, presented below.

Automating the Proposed Methodology by Annotating the Corpus Automatically

As mentioned in this thesis, the only prior knowledge that our methodology uses, is a corpus annotated with instances of the concepts that we want to discover how they are semantically related. We can further automate the proposed methodology by annotating the corpus automatically. We are planning to examine different approaches in order to automatically annotate our corpus with concepts instances.

We are planning to use previous work [20] on unsupervised semantic class induction. Specifically, we will apply our previous work on a domain specific corpus in order to create classes of words[1] which are semantically similar. Then, hoping that the semantic classes will contain words that are instances of concepts that are related with the domain of corpus, we will annotate the corpus. We believe that this approach will work because, the experimental results of this method demonstrate that the produced semantic classes contain similar semantic words with high precision. Another approach for automatically annotate the corpus is by using name entity recognition systems. Nowadays, already exist effective, in terms of precision and recall, systems and methods for terms extraction. We will use these systems in order to extract terms of different concepts and with these extracted terms to annotate our corpus.

Both of the aforementioned approaches for automatic annotation of corpus are mentioned in annotating a corpus with instances of Low-Level concepts (constituted from only one attribute). The automatic annotation of a corpus with instances of High-Level concepts (constituted from more than one attributes), is a very difficult task and as far as we know a reliable approach for this task does not existin the literature. Nevertheless, we are planning firstly to annotate the corpus with instances of Low-Level concepts, apply our algorithm for discovering how these concepts are related. Then, assuming each of the clusters of related concepts, as we saw in section 5.3.2, as a High-Level concept and the related Low-Level concepts as its attributes, we again re-annotate the corpus with instances of High-Level concepts this time. In this manner will try to “build” automatically and unsupervised an ontology from text corpora by

[1] We will try to cluster in semantic classes all words of the corpus except of the verbs, the adjectives, the adverbs and the stop words.

applying our proposed method in a bootstrapping way.

Moreover, in order to overcome the lack of annotated corpora, one could use wikipedia's pages in order to automatically create annotated corpus for a domain. More specifically, having a domain, e.g "Computer Science", you could get wikipedia's page mention on that domain. This page will contains terms which are related with the domain and are also linked to pages that explain these terms. In this manner and repeating for 3-levels links, you could collect a big[2] corpus. Of course this corpus it contains "noisy" terms, which you could first to filtrate with heuristics[3] or algorithms. In this corpus you could either apply one of the aforementioned approaches in order to annotate it or you could use as annotations the terms that contain links, assuming each of these terms as a different concept. Finally, applying our methodology you aim to create for a domain an ontological relational schema of concepts that are related and describe that domain. In other words, given a domain, you could automatically and unsupervised to create an ontology with concepts about this domain.

Characterizing and Labelling the Discovered Semantic Relations

The proposed methodology is able to discover directed semantic relations among concepts. These semantic relations could either be taxonomical or non-taxonomical. The proposed algorithm with the presented methodology is not able to characterize a semantic relationship as taxonomic or non-taxonomic. As mentioned above, our methods in order to discover semantic relations, are based on overlapping instances of different concepts and specifically, in a binary way, overlapping or not overlapping. If we extend this assumption and also enhance it by examining the percentage of the overlapping set. We can characterize the discovered relation as taxonomic(hierarchical) or non-taxonomic(non-hierarchical) and we will also be able to further characterize the taxonomic relations as hyponym or hypernym relations.

If two concepts A and B are related taxonomically and more specific B is hyponym of A (A is hypermyn of B). Then, the sets that represent the offsets of A's instances must be subsets or at least equal to the sets that represent the offsets of B's instances. We claim that because, the concept B as hyponym of A will be constituted from the attributes of A plus some other attributes that further subcategorize concept A. In other words, an instance of B is also an instance of A plus some extra information, because B is further subcategorize concept A. Consequently, the offsets of A's instances are overlapping 100% form the offsets of B's instances, because A's offsets are subsets

[2] If each page contains about 30 links on relevant domains, for a 3-level links you could collect about 1000 texts

[3] A simple heuristic-rule might be, keep only the terms that appeared in at least 3 different pages.

of B's instances. According to the above syllogism, we can extend our methodology in order to characterize the discovered relations as taxonomic or non-taxonomic and for the taxonomic relations as hyponym or hypernym. The experimental results in both domains, presented in sections 5.3.1 and 5.4.1, endorse our above syllogism.

Finally, another aspect for future work is also to label the non-taxonomic semantic relations. For this purpose we are planing to use already existing approaches, e.g. [24], etc. Another approach is to examine the role of the verbs, in a statistical approach, that appear between the instances of the related concepts.

Appendix A

Semantic-Correlation Scores

A.1 Semantic-Correlation Scores between High-Level concepts of Athletic Domain

The Semantic-Correlation Scores $S(RunningCompetition \rightarrow ?)$ sorted, between *"RunningCompetition"* concept and and each of the rest of the concepts for all Boemie corpus:

$$S(RunningCompetition \rightarrow SportsEvent) = 0.386$$

$$S(RunningCompetition \rightarrow SportCompetition) = 0.380$$

$$S(RunningCompetition \rightarrow Athlete) = 0.344$$

$$S(RunningCompetition \rightarrow SportsTrial) = 0.275$$

$$S(RunningCompetition \rightarrow MaleAthlete) = 0.218$$

$$S(RunningCompetition \rightarrow FemaleAthlete) = 0.208$$

$$S(RunningCompetition \rightarrow HurdlingCompetition) = 0.123$$

$$S(RunningCompetition \rightarrow MarathonCompetition) = 0.122$$

$$S(RunningCompetition \rightarrow Running100mCompetition) = 0.111$$

$$S(RunningCompetition \rightarrow RaceWalkingCompetition) = 0.080$$

$$S(RunningCompetition \rightarrow SportsRound) = 0.039$$

$$S(RunningCompetition \rightarrow JumpingCompetition) = 0$$

$$S(RunningCompetition \rightarrow \dots) = 0$$

The Semantic-Correlation Scores $S(RunningCompetition \rightarrow ?)$ sorted, between *"RunningCompetition"* concept and and each of the rest of the concepts for the Boemie corpus, witout containg annotations for the *"MarathonCompetition"* concept:

$S(RunningCompetition \rightarrow SportCompetition) = 0.401$

$S(RunningCompetition \rightarrow SportsEvent) = 0.375$

$S(RunningCompetition \rightarrow Athlete) = 0.354$

$S(RunningCompetition \rightarrow SportsTrial) = 0.269$

$S(RunningCompetition \rightarrow MaleAthlete) = 0.223$

$S(RunningCompetition \rightarrow FemaleAthlete) = 0.220$

$S(RunningCompetition \rightarrow HurdlingCompetition) = 0.182$

$S(RunningCompetition \rightarrow Running100mCompetition) = 0.165$

$S(RunningCompetition \rightarrow RaceWalkingCompetition) = 0.119$

$S(RunningCompetition \rightarrow SportsRound) = 0.058$

$S(RunningCompetition \rightarrow JumpingCompetition) = 0$

$S(RunningCompetition \rightarrow \ldots) = 0$

A.2 Semantic-Correlation Scores between Low-Level concepts of Athletic Domain

The Semantic-Correlation Scores $S(round\_name \rightarrow ?)$ sorted, between *"Round-Name"* concept and and each of the rest of the Low-Level concepts for all Boemie corpus:

$S(round\_name \rightarrow sport\_name) = 0.610$

$S(round\_name \rightarrow gender) = 0.605$

$S(round\_name \rightarrow name) = 0.512$

$S(round\_name \rightarrow nationality) = 0.354$

$S(round\_name \rightarrow ranking) = 0.291$

$S(round\_name \rightarrow date) = 0.257$

$S(round\_name \rightarrow performance) = 0.204$

$S(round\_name \rightarrow event\_name) = 0.078$

$S(round\_name \rightarrow age) = 0.044$

$S(round\_name \rightarrow city) = 0.039$

$S(round\_name \rightarrow country) = 0.013$

$S(round\_name \rightarrow stadium\_name) = 0.008$

Bibliography

[1] Bechhofer, S., Horrock, I., Goble, C., and R., S., "*OilEd: A Reason-able ontology editor for the semantic web.*." In Joint German/Australian Conf. on Articial Intelligence., 2001.

[2] Berland, M. and Charniak, E., "*Finding parts in very large corpora.*." In proceedings of the 37^{th} annual meeting of Association for Computational Linguisitcs., 1999.

[3] Berners-Lee, T., Hendler, J., and Lassila, O., "*The semantic web.*" In Scientic American, 2001.

[4] Caraballo, S. A., "*Automatic construction of hypernym-labeled noun hierarchy from text.*." In proceedings of the 37^{th} annual meeting of the Association for Computational Linguistics., 1999.

[5] Cederberg, S. and Widdows, D., "*Using LSA and noun coordination information to improve the precision and recall of the hyponymy extraction.*." In Conference on Natural Language Learning., 2003.

[6] Church, K., and Hanks, P., "*Word Association Norms, Mutual Information and Lexicography.*." In Computational Linguistics, Vol 16:1, pp. 22-29, 1991.

[7] Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., and Rojas, I., "*Unsupervised learning of semantic relations between concepts of a molecular biology ontology.*." In International Joint Conference on Articial Intelligence., 2005.

[8] D. Lin and P. Pantel., "*Dirt:Discovery of Inference Rules from Text.*." In In ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 323328., 2001.

[9] Daille, B., "*Conceptual structuring through term variations.*." In ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pages 916., 2003.

[10] Espinosa, I.S., Kaya, A., Melzer, S., Moeller, R., "*On Ontology Based Abduction For Text Interpretation..*" In Proceedings of CICLing 2008, pp. 194205, 2008.

[11] Farquhar, A., Fikes, R., and Rice, J., "*The ontolingua server: A tool for collaborative ontology construction..*" International Journal of Human-Computer Studies, 46:707727., 1997.

[12] Faure, D. and Nedellec, C., "*Asium: Learning sub-categorization frmaes and restrictions of selection..*" In 10^{th} European Conference on Machine Learning., 1998.

[13] Fotzo, H. N. and Gallinari, P. , "*Information access via topic hierarchies and thematic annotations from document collections..*" In International Conference on Enterprise Information Systems, pages 6976., 2004.

[14] Girju, R. and Moldovan, D., "*Text mining for causal relations..*" In 15^{th} international Florida Articial Intelligence Research Society Conference, pages 360364., 2002.

[15] Girju, R., Badulescu, A., and Moldovan, D., "*Learning semantic constraints for the automatic discovery of part-whole relations..*" In Human Language Technologies and North Ameircan Association of Computational Linguisitcs, pages 8087., 2003.

[16] Gomes-Perez, A., Fernadez-Lopez, M. and et al., "*Ontological Engineering: with examples from the areas of Knowledge Management, e-commerce and the Semantic Web.*" Book by Springer, 2004.

[17] Gruber, T. R., "*Toward principles for the design of ontologies used for knowledge sharing..*" Journal of Human Computer Studies, 43(5):907928., 1993.

[18] Guarino, N. and Giaretta, P., "*Ontologies and knowledge bases: Towards a terminological clarication..*" In N.Mars(ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. IOS Press, Amsterdam: 25-32, 1995.

[19] Hearst, M., "*Automatic acquisition of hyponyms from large text corpora.*" In 14sup th International Conference on Computational Linguistics., 1992.

[20] Iosif, E., Tegos, A., Pangos, A., Fosler-Lussier, E., Potamianos, A., "*Unsupervised Combination Of Metrics For Semantic Class Induction..*" IEEE Spoken Language Technology Workshop, 2006.

[21] Iwanska, L., Mata, N., and Kruger, K., "*Fully automatic acquisition of taxonomic knowledge from large corpora of texts: Limited syntax knowledge representation system based on natural language..*" In 11^{th} International Symposium on Methodologies for Intelligent Systems, pages 430438., 1999.

[22] Jacquemin, C., "*A symbolic and surgical acquisition of terms through variation.*." In Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, pages 425438., 1996.

[23] Kashyap, V., Ramakrishnan, C., Thomas, C., Bassu, D., Rindesch, T. C., and Sheth, A., "*Taxaminer: An experimental on framework for automated taxonomy bootstrapping.*." Technical report, University of Georgia., 2004.

[24] Kavalec, M., Maedche, A., and Svatek, V., "*Discovery of lexical entries for non-taxonomic relations in ontology learning.*." In SOFSEM., 2004.

[25] Lackoff, G., Johnson, M., "*Metaphors we live by.*" University of Chigago Press, 1980.

[26] Lackoff, G., Johnson, M., "*Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought.*" Basic Books, 1997.

[27] Liakata, M. and Pulman, S., "*Learning theories from text.*" In proceedings of COLING 18, Geneva, Switzerland., 2004.

[28] Mädche, A. and Staab, S., "*Ontology Learning for the Semantic Web.*" IEEE Intelligent Systems, 16(2):72-79., 2001.

[29] Maedche, A. and Volz, R., "*The ontology extraction and maintenance framework text-to-onto.*." In International Conference on Data Mining., 2001.

[30] Morin, E. and Jacquemin, C., "*Automatic acquisition and expansion of hypernym links.*." In Computer and Humanities., 2003.

[31] Petasis, G., Karkaletsis, V., Paliouras, G., Spyropoulos, C., "*Learning context-free grammars to extract relations from text.*." In Proceedings of ECAI-2008, pp. 303-307, 2008.

[32] Quan, T. T., Hui, S. C., Fong, A. C. M., and Cao, T. H., "*Automatic generation of ontology for scholarly semantic web.*." In International Semantic Web Conference., 2004.

[33] Quinlan, R. J., "*C4.5: Programs for Machine Learning.*." Morgan Kaufmann., 1993.

[34] Robert-Alain de Bengrade, Dressler, W., "*Introduction to text Linguistics.*" Longman, August 24, 1981.

[35] Ryu, P. and Choi, K. S., "*Measuring the specicity of terms for automatic hierarchy construction.*." In European Conference on Articial Intelligence Workshop on Ontology Learning and Population., 2004.

[36] Sabou, M., Wroe, C., and Goble, C., "*Learning domain ontologies for web service descriptions: an experiment in bioinformatics.*." In 14th International World Wide Web Conference., 2005.

[37] Sahlgren, M., "*The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.*" Phd Thesis, Department of Linguistics-Computational Linguistics, Stockholm University, 2006.

[38] Schutz, A. and Buitelaar, P., "*Relext: A tool for relation extraction from text in ontology extension.*." In Fourth International Semantic Web Conference., 2005.

[39] Schutze, H., "*Word Space.*" In: Proc. Conference on Advances in Neural Information Processing Systems (NIPS), 1993.

[40] Shamsfad, M. and Barforoush, A., B., "*Learning ontologies from natural language texts.*." International Journal of Human-Computer Studies, 60(1):1763., 2003.

[41] Snow, R., Jurafsky, D., and Ng, Y. A., "*Learning syntactic patterns for automatic hypernym discovery.*." In Advances in Neural information Processing Systems., 2004.

[42] Staab, S. and Studer, R., "*Handbook on Ontologies.*" International Handbboks on Information Systems. Springer, 2004.

[43] Stanford Center for Biomedical Informatics Research, "*Protege.*" http://protege.stanford.edu. Technical report, 2001.

[44] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., and Wenke, D., "*Ontoedit: Collaboratve ontology development for the semantic web.*." In International Semantic Web Conference., 2002.

[45] Turney, P. D., "*Expressing implicit semantic relations without supervision.*." In proceedings of the annual meeting association for computational linguistics (ACL-2006), pages 313320., 2006.

[46] Valarakos, A., Karkaletsis, V., Alexopoulou, D., Papadimitriou, Spyropoulos, C., Vouros, G., "*Building an Allergens Ontology and Maintaining it using Machine Learning Techniques.*." In Computers in Biology and Medicine Journal (CBM), 36 (10): 1155-1184, 2006.

[47] Yates, B. R. and Neto, R. B., "." Modern Information Retrieval. Addison Wesley., 1999.

[48] Zavitsanos, E., Paliouras, G., Vouros, G., "*Ontology Learning and Evaluation: A survey..*" Technical Report, NCSR Demokritos DEMO-2006-3, 2006.