• U  C •

MSc Thesis
in
Informatics Engineering

Automatic Extraction and Validation of Lexical Ontologies from text

Hernani Pereira Gomes Costa

Thesis Advisors:
Hugo Gonçalo Oliveira
Paulo Gomes

Department of Informatics Engineering
Faculty of Sciences and Technology
University of Coimbra

September 2010

Acknowledgements

I would like to express gratitude to Durval Pires, Carlos Oliveira and especially to Bárbara Furtado for the care with which they reviewed this thesis; and for conversations that clarified my thinking on this and other matters. Their friendship and professional collaboration meant a great deal to me. I am indebted to the *KDigg* group, that demonstrated disponibility to review the results created in this thesis, made comments that encouraged me to revise and improve this research. I am particularly grateful to Hugo Gonçalo Oliveira for his thoughtful, creative comments, time shared and more generally for exploring with me the boundaries of professional friendship.

This thesis was developed from ideas published in Hearst (1992), Oliveira (2009) MSc thesis, and particularly in Gonçalo Oliveira (2009) PhD proposal. Finally, I must acknowledge the *Knowledge and Intelligent System Laboratory*, for allowing me to use their resources.

As always, it was Paulo Jorge de Sousa Gomes who provided the shelter conditions under which the work could take place - thanks to him for this and many other things.

Glossary

AC/DC	Acesso a Corpos/ Disponibilização de Corpos
AI	Artificial Intelligence
IE	Information Extraction
IR	Information Retrieval
MRDs	Machine Readable Dictionaries
MT	Machine Translation
NER	Named Entity Recognition
NLP	Natural Language Processing
OWL	Web Ontology Language
POS	Part of Speech
QA	Question Answering
RDF	Resource Definition Framework
WSD	Word Sense Disambiguation

Abstract

Nowadays, semantic information plays an important role in natural language processing, more specifically describing and representing "the meanings of the words" crucial for understanding the human language.

In the last two decades, there have been efforts to create a large database that represents lexical knowledge, where the words and their meanings are represented along with connections held between them. However, in most of the cases, this resources are created manually. For instance Princeton WordNet is considered the standard model of a lexical ontology for the English language. Besides that, also for Portuguese there have been some attempts to create a broad-coverage ontology, also created manually and not publicly available. Still, they are not public available for download, and also all of them were manually created. Despite being less prone to errors, the problem is that the manual creation of these resources takes a lot of time consuming and requires a team, and researchers specialised in the area.

Nevertheless, in the last years, some efforts have been made to develop computational tools to reduce the need of manual intervention, such as some authors that propose lexico-semantic patterns to find semantic relations between terms in text. This kind of approach should be considered as an alternative and subject of research, in order to avoid impractical human work in the construction of these resources.

Having this in mind, the work expected in this project is the creation of a system capable of automatically acquire semantic knowledge from any kind of Portuguese text. The extraction method is based on lexico-syntactic patterns, that indicate a relation of interest, and also by a inference method to extract hypernymy relations from compound nouns. Also, different kinds of textual resources are used to test and improve our system.

Furthermore, this work analyses the benefits from applying similarity distributional metrics based on the occurrence of words in documents to our system outputs.

The quality and the utility of the knowledge extracted from the various textual resources, will be compared against another Portuguese knowledge-base.

In the end of this research, important contributions for the computational processing of Portuguese language are provided, such as computational tools capable of extracting and inferring lexico-semantic information from text, methodologies to automatically validate these knowledge, and also compare knowledge-bases. Finally, the experiments outcomes and conclusions are published in important conferences for the area.

Keywords: information extraction, information retrieval, lexical ontologies, lexico-syntactic patterns, semantic knowledge, semantic relations.

Resumo

Hoje em dia, a informação semântica tem um papel muito importante no processamento de linguagem natural, mais especificamente na descrição e representação dos "sentidos das palavra", crucial para a compreensão da linguagem natural.

Nas duas últimas décadas, têm sido feitos esforços no sentido de criar uma ontologia de larga cobertura que represente conhecimento lexical, onde as palavras e os seus significados são representados através de relações entre si. No entanto, na maioria dos casos, estes recursos são criados manualmente. Como por exemplo o WordNet de Princeton, considerado o modelo standard de uma ontologia lexical para a língua Inglesa. Também para o Português têm existido algumas tentativas na criação de uma ontologia de larga cobertura da língua, mas também são criados de modo manual e não estão disponíveis ao público. Apesar de serem menos propícios a erros, estes recursos demoram imenso tempo a serem criados e, para isso, requerem uma equipa de investigadores especializados na área.

Todavia, nos últimos anos, tem existido um grande empenho em desenvolver ferramentas computacionais que reduzam a necessidade de intervenção manual. Assim, alguns autores propõem o uso de padrões léxico-sintácticos em texto, para procurar relações semânticas entre termos.

Tendo isto presente, esperamos com este projecto criar um sistema capaz de obter automaticamente conhecimento semântico a partir de qualquer tipo de texto em Português. O método de extracção é baseado em padrões lexico-sintácticos que indiquem uma relação de interesse; e também através de um método de inferência para extrair relações de hiperonímia de termos compostos. São também usados diferentes tipos de recursos textuais para testar e melhorar o nosso sistema.

Além disso, este trabalho analisa os benefícios da aplicação de métricas de semelhança distribucionais, baseadas na ocorrência de palavras nos documentos, no conhecimento gerado pelo nosso sistema.

A qualidade e a utilidade do conhecimento extraído pelo nosso sistema nos vários recursos textuais, será depois comparado com outra base de conhecimento em Português.

No final desta investigação, são fornecidas importantes contribuições para o processamento computacional da língua Portuguesa, assim como: ferramentas computacionais capazes de extrair e inferir conhecimento lexico-semântico a partir de texto; metodologias para validar automaticamente esse conhecimento; e também comparar bases de conhecimento. Finalmente, os resultados das experiências e as suas conclusões são publicadas em importantes conferências da área.

Palavras-Chave: conhecimento semântico, extracção de informação, ontologias lexicais, padrões léxico-sintácticos, recuperação de informação, relações semânticas

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays we live in a world that is surrounded by information, most of the times provided as natural language text. In order to exploit this written data, many applications are being developed for performing different tasks where understanding the meaning of natural language is critical. Knowledge management (Gaines and Shaw (1997)), exchange of electronic information (Boss and Ritter (1993)) or the Semantic Web (Berners-Lee et al. (2001)) are just some of the areas where we can see this kind of applications. They demonstrate that natural language processing (NLP) (Jurafsky and Martin (2000)), has become more and more dependent on semantic information and so, computational access to such type of knowledge is important and some times indispensable.

For making people and machines communicate in the same language, it is necessary to develop tools capable of exchanging well-defined and unambiguous information. Therefore, it is crucial that tools are able to manipulate natural language and to encode it into a formal language, interpretable unambiguously by machines. Lexical databases, lexical knowledge-bases or lexical ontologies are some of the names given to the resources resulting from these efforts. Since the interpretation of meaning is entirely linked to the knowledge of those who communicate and there is not a simple method that allows us to formalise all the information that the humans share between them, this area is quite challenging. There have been several attempts to extract lexico-semantic information from written data, such as Hearst (1992), Baségio (2007) and Freitas (2007), that propose some lexico-semantic patterns to find semantic relations between terms in text.

Besides that, there have been some attempts to formalise semantic knowledge in a kind of lexical ontology, such as Princeton WordNet (Fellbaum (1998)), considered as a reference model for the English language. Similar resources for Portuguese are being created (WordNet.BR (Dias-da-Silva (2006)), WordNet.PT (Marrafa et al. (2006))), however they are not publicly available for download. Moreover, all of them, were handcrafted.

However, as many authors have shown (Gonçalo Oliveira et al. (2010b) (PAPEL), Gonçalo Oliveira and Gomes (2010) (Onto.PT)), taking advantage of available NLP tools, it is possible to create a system capable to automatically extract semantic knowledge from text, reducing the need of manual intervention. These approaches should be considered as an alternative and a subject of research, in order to avoid time-consuming effort and impractical human work in the construction of these resources.

Nevertheless, when referring to text that is not associated with a specific domain, it is more difficult to anticipate what kind of information can be found and plausible to extract. However, instead of analysing all the text, we can only seek the most relevant information, which may be discovered through a simple understanding of the text, i.e., through a linguistic analysis. We are aware that some noise can result from this automatic approach.

For instance, consider the following sentence:

A car is a vehicle that has an engine and aims to move planets.[1]

A simple algorithm could be used to find out that *car* is-a *vehicle*, *car* has-a *engine* and *car* purpose_of *move planets*, using only three lexical patterns that indicate the semantic relations, such as "*is a*", "*has*" and "*aims to*". Nonetheless, there are many ways that the structure of a language can indicate the meaning of lexical items, and the difficulty lies in finding constructions that reliably indicate the relation of interest. Still, in the literature (Girju et al. (2006), Khoo et al. (2000), Girju and Moldovan (2002), among other, we can find some lexico-syntactic patterns that frequently indicate the relation of interest and occur frequently in many text genres. Despite that, it is known that only a subset of the possible instances of a relation of interest will appear in a particular form. So, following Hearst (1992), we need to automatically discover new patterns that indicate a relation of interest and make use of them.

Furthermore, there is a problem with the semantics of the sentence above, which would not be detected with the pattern-based algorithm or even with the new discovered patterns. However, if we improve this algorithm with an additional statistical module, it would eventually understand that the words *car* and *move planets* do not co-occur frequently in text, and so the apparent relations between *car* and *move planets* could be ignored. We believe that an interesting approach to deal with the limitations of systems capable of acquiring semantic knowledge from text, including the aforementioned, would be to weight their outputs according to the occurrences of words in text, creating that way a hybrid system.

Having this in mind, the first objective for this research is to perform an automatic discovery of new patterns, like Hearst (Hearst (1992)) proposed. Another objective is the creation of a hybrid system capable of automatically extracting semantic knowledge from any kind of unstructured text, such as documents, corpus or even Web files, and also infer new knowledge based on compound terms[2]. This extraction system belongs to Onto.PT[3] (Gonçalo Oliveira and Gomes (2010)) in whose this work is integrated.

To do that, besides the creation of a module capable of extract semantic knowledge from text, it is necessary to create a module that extracts written data from any kind of textual resource. Also, it is necessary a module capable of, in a automatic way, validate this knowledge. It considers the entities co-occurrence in corpus to verify if these entities are really related in a semantic way. Last but not least, using the knowledge extracted from our system combined with indicative textual patterns,

[1] In Portuguese: *Um automóvel é um veículo que tem um motor e tem como finalidade transportar planetas.*

[2] Compound terms are built by combining two (or more) simple terms.

[3] http://ontopt.dei.uc.pt

we propose a method to, in an automatic way, evaluate semantic knowledge in the Web.

In order to test and improve our system, different kinds of textual corpora will be exploited. The knowledge extracted from these resources will be then compared to another lexical resource (PAPEL (Gonçalo Oliveira et al. (2010b))) in order to study there commonest and completeness.

Also, we use human judges to evaluate and verify the reliability of the extraction and validation processes.

At the end of this research, the following contributions are expected:

- Computational tools for automatic:
  - discovery of new semantic patterns;
  - extraction and inference of semantic knowledge from textual resources;
  - validate and evaluate semantic knowledge;
  - comparison of knowledge-bases.
- Scientific papers about the most relevant conclusions and results from the experiments carried out.
- MSc thesis, describing all the work done in the research.

The outline of this MSc thesis consists of four more chapters:

Chapter 2 introduces the background concepts, important to understand this research. This includes: an introduction to natural language processing, section 2.1; a section that covers concepts related to ontologies, such as their definition, categorisation, construction, and their real applications, section 2.2; an overview on related work is made in section 2.3; an overview of some linguistic resources that can be used to test and make certain decisions related to our work, are presented in section 2.4; some similarity measures used to validate data produced by our system, are presented in section 2.5; some of the tools and libraries that will be used in this research are described in section 2.6; finally, in section 2.7, a summary of this chapter is presented.

Chapter 3 explains and presents all the modules of our system, including their expected outcomes. More specifically: in section 3.1 the module responsible for extracting written data from textual resources, including the inference approach, is presented; section 3.2 describes the module which automatically extracts semantic knowledge from text; and in section 3.3 it is presented a module capable of quantify this knowledge. The last section of this chapter, section 3.4, proposes a method to automatically compare different knowledge-bases.

Chapter 4 reports the results of four experiments carried out: the first experiment is an experimental approach in CETEMPúblico, using a simple version of our system, section 4.1; the second experiment is an approach in the Wikipedia abstracts, a free text corpus, using a better version of the system, section 4.2; the third experiment describes a second approach in the CETEMPúblico corpus, comparing

the first versus the second version of our system, section 4.3; the fourth experiment, and the last one, presents an experimental approach, that analyses the quantity of common knowledge between three resources, section 4.4.

Chapter 5 describes a summary of this thesis, discusses its contributions, presents the resulting publications, and provides ideas for future research.

Chapter 2

Background Knowledge

This chapter provides the most relevant information for understanding this research. It starts with a short introduction to natural language processing (NLP) (section 2.1), including some levels related and involved in this research, followed by a description of well-know NLP tasks (section 2.1.4), like information extraction (IE) and information retrieval (IR).

Fundamental concepts related to ontologies is presented in section 2.2, such as their definition, categorisation, construction, and their real application. This kind of resources are very important for the knowledge representation. Then, an overview on related work is made in section 2.3. In the section 2.4 some linguistic resources relevant to this research are presented.

Besides the NLP concepts, NLP tasks, manners to represent knowledge, similar works and resources important to our research, we will explore some distributional metrics, section 2.5, typically used to validate data. Further, some of the tools and libraries that will be used in this system are described in section 2.6.

Finally, in the last section of this chapter, section 2.7 presents a summary of this chapter. It explains the connection between all the aforementioned sections, relating them and explaining our research approach and purpose.

2.1 Natural Language Processing

Natural language processing (NLP) is a very important field of artificial intelligence (AI) and linguistics that aims to develop techniques that allow to generate and automatically understand the language that humans use to communicate among themselves. It is an area under development since the 1940s, however there is still much to do.

Computers are able to interpret written instructions through their own language (formal language) which uses fixed rules and well defined logical structures, however humans use a different type of language to communicate, natural language. The greatest difficulty in processing natural language into formal language is its ambiguity, with can occur in several levels: phonology, morphology, syntax, semantics and pragmatics. In this work, we focus on the three most relevant levels to this thesis (morphological, syntactic and semantic), which are introduced in the next subsections.

2.1.1 Morphological Level

Morphology is the branch of linguistics that studies the internal structure of words, in order to identify, analyse and describe the 'word' itself. At this level words are studied independently, without taking into account the structure of the sentence where, and the order in which they occur. Morphological analysis consists in determining each word's morphological category (or categories), identifying its base form (called lemma), and also other features that depend on its category, such as its gender, number or tense. For nouns, the lemma is usually a word in the masculine genre and singular number, while for verbs it is a word in the infinity form. Two simple examples of morphological analysis may suffice to illustrate this analysis.

- as palavras: *professor* (singular, masculino), *professores* (plural, masculino) e *professoras* (plural, feminino), têm o mesmo lema, *professor*.
- as palavras *fui* (1ªpessoa do singular no Pretérito Perfeito) e *serei* (1ªpessoa do singular no Futuro) têm como lema *ser*.
- the words: *teacher* (singular, male), *teachers* (plural, masculine) and *teachers* (plural, feminine), have the same lemma, *teacher*.
- the word *was* (1st person singular past tense) and *will be* (1st person singular Future), have the same lemma *be*.

The following examples shows the kind of ambiguity that may occur at morphological level:

O Zé ***casa*** com a Bárbara.	Zé ***marries*** Bárbara.
O Zé irá para ***casa***.	Zé will go ***home***.

In the first example the word *casa* is a form of the verb *casa* in the 3^{rd} person singular in the simple present. But in the second example we have the same word, but this time it is used as a noun. When performing a morphological analysis in the presented sentences it is not possible to assign a unique grammatical category to the word *casa*, because this word can occur as a noun or as verb. To identify the exact category of the word in its context, a syntactic analysis is needed to take in consideration the word's neighbourhood and the structure in which it is inserted.

2.1.2 Syntactic Level

Syntax studies the structural relationships between words within the same sentence. Sentences follow a finite set of rules and principles allowing writers and readers to recognize their significance.

Words can be grouped according to their functions. Depending on their place in the sentence and the words around them, they can have different grammatical categories, commonly referred as part of speech (POS), that are usually one of the possible morphological categories for the word. Using algorithms that determine the part-of-speech (POS) of the word (e.g. Chomsky (1956)), it is possible to determine their function in the sentence.

Using computational modules (like: grammars, algorithms that analyse the relationships between words, statistical modules, syntactic trees) morphology disambiguation, which could not be performed in the previous level, will be possible. Moreover, the POS give us information about how the word is pronounced, and it can be very helpful in other NLP tasks, such as information retrieval (IR) or

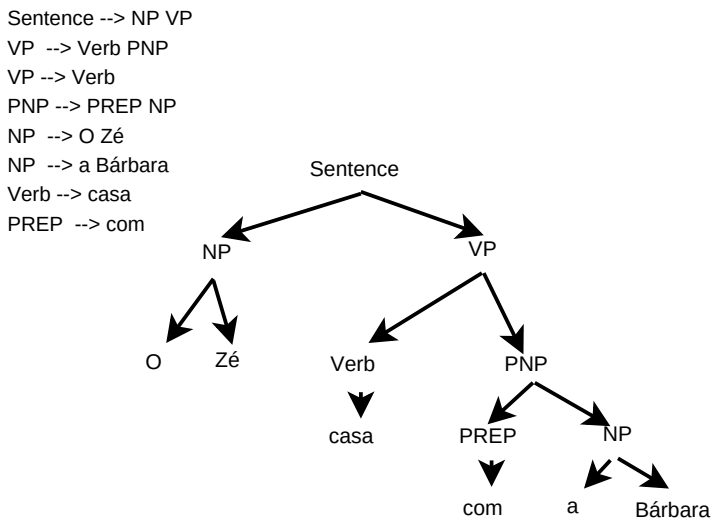


Figure 2.1: Parse tree example.

word sense disambiguation (WSD). For example, given a context-free grammar a sentence can be derived. Figure 2.1 shows the grammar and its derivation tree for the following sentence: *O Zé casa com a Bárbara* (see table 2.1).

However, ambiguities in the syntactic analyses may occur in distinct situations, such as the following sentence:

S(1): *O homem viu a rapariga com o telescópio.*[1]

- *O*/artigo *Zé*/nome *casa*/verbo *com*/preposição *a*/determinante *Bárbara*/nome.
- *O*/artigo *Zé*/nome *ir*/verbo *para*/preposição *casa*/nome.
- *Zé*/noun *marries*/verb *Bárbara*/noun.
- *Zé*/noun *will go*/verb *home*/noun.

Table 2.1: Syntactic analysis.

This sentence, S(1), raises the following questions: - Who has the telescope? - The man or the girl?

Even for humans is hard to tell who has the telescope without other elements that help to understand the context.

2.1.3 Semantic Level

As we have seen before, syntax studies the rules and principles on how to create syntactic expressions that can be interpreted from simpler expressions, however by itself does not assign meanings, hence the origin of semantics, which is designed to study the sense of language (syntactically well formed). Nevertheless, computers can not interpret this kind of knowledge (sense of a word, phrase, judgement or even collections of text) because they need to relate natural language with formal language.

There are several ways to represent semantic into formal language, for example by logical predicates (Smullyan (1995)), directed graphs or semantic frames (Fillmore

[1] In English, *The man saw a girl with the telescope.*

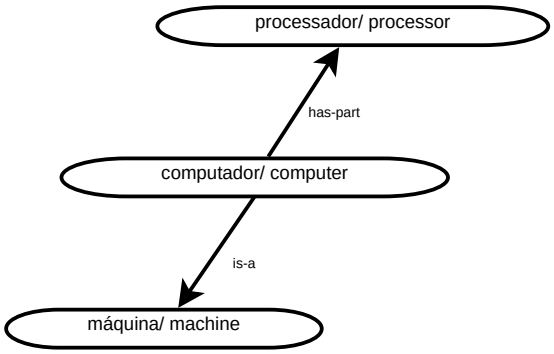


Figure 2.2: Representation of the meaning of the phrase by a directed graph.

(1982)). Tables 2.2, 2.3 and figure 2.2 presents the aforementioned representation of the meaning of sentence S(2).

S(2): *Um computador é uma máquina que tem um processador.*[2]

é-uma (computador, máquina)
tem-parte (computador, processador)
isa (computer, machine)
has-part (computer, processor)

Table 2.2: Representation of the meaning of the phrase by logical predicates.

computador
é-uma: máquina
tem-parte: processador
computer
isa: machine
has-part: processor

Table 2.3: Representation of the meaning of the phrase by the *frame computer*.

In both representations there are connections between the meaning of represented words, commonly called as semantic relations. In this document, semantic relation instances will be represented by: *entity_1 RELATION_NAME entity_2*, with the relation name in portuguese. As we can see in table 2.4, this is very similar to the logical predicates.

máquina ***HIPERONIMO_DE*** *computador*	*machine* ***HYPERNYM_OF*** *computer*
computador ***HOLONIMO_OF*** *processador*	*computer* ***HOLONYM_OF*** *processor*

Table 2.4: Representation of the meaning of the phrase by semantic relations.

Next will be presented some more examples of hypernymy relation, with a brief explanation, as well as other semantic relations with relevance to this work.

[2]In English, *A computer is a machine that has a processor.*

Hyponymy and hypernymy relationship

Hypernymy is the most known and studied semantic relations, which relates concepts[3]. An entity Y is a hyponym of an entity X if Y is a subtype or instance of X. For example:

calçado ***HIPERONIMO_DE*** *sapatos*	*footwear* ***HYPERNYM_OF*** *shoes*

The word *"footwear"* is more general concept (hypernym), while the word *"shoe"* is a more specific concept (hyponym), which is a kind of *"footwear"*.

melro ***HIPONIMO_DE*** *ave*	*blackbird* ***HYPONYM_OF*** *bird*
pardal ***HIPONIMO_OF*** *ave*	*sparrow* ***HYPONYM_OF*** *bird*

In this example, the words *"sparrow"* and *"blackbirds"* are identified as a hyponyms of *"bird"* because they represent types of species, in this case birds, and birds is the hypernym of the sentence.

Holonymy and meronymy relation

Meronymy, also known as part-of is another type of semantic relationship. Holonymy and meronymy relate concepts, such that one belongs, is part, piece or member of another.

roda ***PART_OF*** *bicicleta*	*wheel* ***PART_OF*** *bicycle*

Holonymy occurs when an entity denotes a whole or "has/ includes" another entity.

bicicleta ***HOLONIMO_DE*** *roda*	*bicycle* ***HOLONYM_OF*** *wheel*

Synonymy relation

Synonymy is a type of semantic relationship between words. Synonyms are words that have identical or very similar meaning, for example:

carro ***SINONIMO_DE*** *automóvel*	*car* ***SYNONYM_OF*** *automobile*
sapatilha ***SINONIMO_DE*** *ténis*	*slippers* ***SYNONYM_OF*** *tennis shoe*

Synonyms are words that belong to the same grammatical category, with similar meaning. Two synonyms may be replaced by one another without changing the meaning of the sentence in which they occur (definition of interchangeability).

Other relationships

Causation: Is a relation between two events (the cause and a the effect).

In the next example the effect is the concept *death* and the cause will be the *poison*.

[3]A concept is a cognitive unit of "meaning" and can contain one or more words.

veneno ***CAUSADOR_DE*** *morte*	*poison* ***CAUSATION_OF*** *death*

Location: When a concept is located in another, for instance:

Coimbra ***LOCALIZADA_EM*** *Portugal*	*Coimbra* ***LOCATED_IN*** *Portugal*

Manner: When an action can be performed in some manner, for example:

rápido ***MANEIRA_DE*** *correr*	*fast* ***MANNER_OF*** *run*

2.1.4 Natural Language Processing Tasks

Combining the aforementioned levels of NLP for different purposes, more complex tasks can be performed, such as the following:

- **Question Answering (QA)** (Strzalkowski and Harabagiu (2006))**:** This task is responsible for giving automatically one or multiple answer(s), to a question in natural language.
- **Machine Translation (MT)** (Hutchins and Somers (1992))**:** automatic translation of text written in one natural language to another.
- **Information Retrieval (IR)** (Salton and McGill (1986))**:** a task concerned with locating documents, other natural languages resources or information within them, according to a user's query.
- **Named Entity Recognition (NER)** (Chinchor and Robinson (1997))**:** recognition, and sometimes classification, of proper nouns (e.g. person's name, organisations, places, events, etc.).
- **Word Sense Disambiguation (WSD)** (Ide and Véronis (1998))**:** task responsible for the selection of the most adequate sense of a word in a context.
- **Information Extraction (IE)** (Grishman (1997))**:** a generic task that has the intention of automatically extracting structured information from unstructured natural language text. IE includes other tasks of NLP tasks, such as NER, relation detection, temporal analysis, etc. There are many ways to extract information in a manner that other applications can process, see for instance Jackson and Moulinier (2002) which provides a method for extracting information:
  - **Tokenization:** this process consists on splitting a body of text into the units of text that take part in the sentences. These tokens can be seen as words, or sequence of characters present in a document, split by a white space.

- **Part of Speech (POS):** using the previous tokens, this process consists on labelling each one with the most adequate POS tag: determinant, adjective, verb, etc.
- **Lemmatisation:** words in natural language appear in different forms, according to there morphological category. Lemmas are the base form, or canonical form of these words. For example, the words 'were', 'is' and 'was' through lemmatisation would be interpreted by their canonical form 'be'.

2.2 Ontologies

In order to share knowledge between people and computers it is essential that the information they want to share is well defined. Building them involves a slow and complex process, which has hampered its use in large scale. This chapter seeks to define the term "ontology", how it can be created, categorised and applied.

2.2.1 Definition

There are various definitions of ontologies, from a simple taxonomy, to a strongly semantical and relational view of domain. The concept of ontology goes back to philosophy, where the 'word' is associated with perception of the nature of the reality around us, linking mind and matter, substance and attribute, fact and value (i.e. the idea is to describe the basic categories and relationships of being or existence, studying and analysing the conception of reality that surrounds us). Another definition that philosophy gives is that the ontology can be seen as the theory of objects and their conventions, defining criteria to distinguish the various objects, both concrete and abstract (existing or not, real or even ideas), and their relationships, dependencies and assertions (claims or arguments).

In the last two decades, the term ontology has been added to the vocabulary of artificial intelligence, which uses ontologies in the process of formal description of the "things of the world", the fundamental process for intelligent systems to consider and act upon world in which they propose to operate (Welty and Guarino (2001)). A definition cited by many is given by Gruber (1993): *"An ontology is an explicit specification of a conceptualisation."*. One last reference (Guarino (1998)), says that are several fields where the importance of ontologies has been recognized, such as knowledge management and organisation of language engineering, information modelling and integration; recovery and extraction of information, database, etc., are just some of the fields where the importance of using ontologies is already recognized.

However, there is still currently a great controversy regarding the definition of ontology, each 'area' has its own definition.

2.2.2 Categorisation

Ontologies must necessarily include a vocabulary of terms and definitions for their meaning, which can differ substantially from the formalism of their definitions.

Uschold and Grüninger (1996) defined an evaluation scale of formalism for ontologies, dividing it into four levels of formality:

- **Highly informal:** the ontology is expressed in unstructured natural language;
- **Semi-informal:** the ontology is expressed in structured and restricted natural language, increasing the clarity of concepts and reducing ambiguity;
- **Semi-formal:** the ontology is expressed in a formally defined artificial language;
- **Rigorously formal:** meticulously defined terms with formal semantics, theorems and proofs of such properties as soundness and completeness.

In terms of content and nature of the concepts present in the ontologies, they can be classified as follow (Guarino and Giaretta (1995)):

- **Application Ontologies** describe concepts that depend on the domain, and tasks related to a specific problem of that same domain (e.g. identifying brain disorders);
- **Central Ontologies or Generic Domain Ontologies:** describe the branches of study in a particular area (e.g. allowed behaviour ontology). The goal is to serve as a base to more specific domain ontologies (e.g. tax law, family, etc.);
- **Domain Ontologies** describe a particular domain in a specific generic area of knowledge (e.g. physiotherapy);
- **Representative Ontologies:** define concepts underlying the formalisms of knowledge representation, setting the representation rules;
- **Generic Ontologies** describe the general definitions of abstract concepts such as: time, space, beings, things, etc., regardless of domain or problem. Necessary to understand the aspects of the world;
- **Task Ontologies** describe concepts and vocabulary related with activities and tasks used in the resolution of problems (e.g. plans, processes, etc.).

It is generally perceived the relationship between the various categories of ontologies; the most general levels are generally reused in the construction of more specific ontologies.

2.2.3 Applications

Uschold and Grüninger (1996) defined that ontologies have three basic types of applications: communications, systems engineering and inter-operability.

Communication

On systems where several agents communicate it is essential that communication between them is possible through a shared interpretation. An ontology can provide a normative model of organisation of information/ knowledge, thereby supporting the integration and communication between the various participants, providing a standardised terminology of objects and relationships in their domains.

Ontologies can be used as an unified structure that allows to reduce the noise of the terminology's organisation, concepts and knowledge sharing, thus promoting communication and knowledge sharing between people from different areas (e.g. Engineering, Philosophy, Linguistics, etc.).

Systems Engineering

Ontologies can support the design and development of software systems. They can assist in the process of identifying system requirements and the various underlying components, i.e., they can help to infer information about the system. for instance, see the CommonKADS methodology (Schreiber et al. (1994)) that use the language Conceptual Modelling Language (CML) to build domain ontologies and tasks in order to support the specification of knowledge-based systems.

Inter-Operability

Ontologies can also be used on environments of different software tools integration, architectures and multi-agent cooperative modules. As these environments involve activities between them, the ontologies in such cases serve as a language of inter-operation, allowing terminology standardisation between them. An example of its application is the project Process Specification Language (PSL) (Schlenoff et al. (1999)) in which the ontologies serve as a translator between two applications, eliminating ambiguities in the definitions of terms used in each system.

2.2.4 Construction

There is not a perfect formula for building an ontology. The process always depends on the purpose. Despite the various methodologies proposed, the most relevant are presented by Uschold and Grüninger (1996) and Noy and McGuinness (2001).

According to Uschold and Grüninger (1996), the construction of ontologies is divided into five processes: three related to the development and two of support, running simultaneously with the first three. Those five processes are described as follows:

1. **Identify purpose and scope:** first it is necessary to define the goals of their construction and their intended applications. The purpose is defined by "purpose issues", e.g. questions in natural language that describe the requirements that must be answered by the ontology.

2. **Building the ontology:** the construction consists of three activities:

   a) ontology capture, i.e., identification of key concepts and relationships in the domain of interest. Production of precise unambiguous text description

for such concepts and relationships. Identification of terms to refer to such concepts and relationships and finally. Agreeing on all of the above;

b) ontology coding, i.e., commit the basic terms that will be used to specify the ontology (e.g. class, entity, relation), choose a representation language (e.g. RDF, OWL) and finally write the code;

c) integrating existing ontologies with existing concepts.

3. **Evaluation:** using "purpose issues", at this moment already in formal language, verify the expressiveness and consistency of the ontology created.

4. **Documentation:** the rules are defined in order to document always aware of the development, to ensure that assumptions used during construction are recorded.

5. **Guidelines for each phase:** define the methods and techniques that will be used as a base for all subsequent phases of construction (e.g. order of execution, input/ output).

2.2.5 Lexical Ontologies

An ontology can be seen as a set of objects or ideas of the world, that relate objects with each other through certain relations (see section 2.1.3). The definition of Lexical Ontology is not consensual, however in Gonçalo Oliveira et al. (2010b), a lexical ontology is defined as a knowledge structure that relates lexical items, in a particular language, between each other through relationship, connecting their meaning. He also refers that it can be seen as a structure that embraces all language, and not just knowledge of a particular domain. Similar to last definition, Wandmacher et al. (2007) define a lexical ontology with an intent to structure information on words of a given language and their semantic relatedness, were the meaning is encoded by relating lexical items between them.

A lexicon (Hirst (2004)) is a list of words in a language, along with some knowledge of how each word is used. Each word, or set of words, which form a lexicon, is described as lexical entries. Entries depend on the purpose of the lexicon and, in particular, may include any of its properties: spelling and phonetic, grammatical category, meaning or use and nature of its relations with other words. It can be seen as an index which maps how to write a word, for the information that describes the world.

A lexical ontology is not a classic ontology, and can be seen as a lexicon that embraces all the language, where words are listed according to their meaning. Despite some people refer to a lexical ontology as an ontology, where the nodes are represented by words and semantic relationships connectors, there is no consensus as regarding its definition (Hirst (2004)).

2.3 Related Work

In this section, we present two approaches directly related to this thesis. These approaches can be used to build new semantic resources or to enrich existing ones.

Most of the related work presented here have in common the fact that they use textual patterns, which indicate semantic relations. Bellow an example is presented:

O melro ***é uma*** *ave.*	*The blackbird* ***is a*** *bird.*
As aves ***possuem*** *um bico.*	*The birds* ***have*** *a beak.*

By taking advantage of the use of the textual patterns *"is a"* and *"have"*, it is possible to identify the following relationships:

ave ***HIPERONIMO_DE*** *melro*	*bird* ***HYPERNYM_OF*** *blackbird*
ave ***HOLONIMO_OF*** *bico*	*bird* ***HOLONYM_OF*** *beak*

The following knowledge extraction approaches are presented separately: the first one extracts information from dictionary, the second uses free text, usually called textual corpora.

2.3.1 Extraction of Semantic Knowledge from Electronic Dictionaries

Machine Readable Dictionaries (MRDs) are electronic versions of common dictionaries, especially designed to be used by computers, normally stored in a database for easy software access, that interpret and manage them (usually through an interface).

Merry Webster's Pocket Dictionary (MPD) and Webster's Seventh New Collegiate Dictionary (W7) are known because they were the first dictionaries in machine-readable format, that were typed manually and distributed on magnetic tape, in the 1960s (Olney et al. (1967)). From that moment, electronic dictionaries began to be seen as a very important resource in the processing of natural languages, and be used as a source of lexical information in the construction of lexical knowledge bases.

A MRD can have additional resources (e.g. language detection, translation, etc.), and because of that it can be called a smart dictionary. Longman Dictionary of Contemporary English (LDOCE)[4] is probably the most representative MRD of the English language. It was developed in the 1980s with the objective of realising the relevance of MRDs for performing NLP tasks (Michiels et al. (1980)).

The importance of using MRDs as a source of lexical information in the various NLP tasks, mainly in the construction of lexical knowledge bases was soon recognised. This is due not only to the fact that dictionaries are restricted in terms of vocabulary (using simple sentences), but mainly because dictionaries are very well structured (i.e., words and there descriptions), making it the main source of lexical knowledge of a language.

The structure of dictionaries, as well as its predictability and the simplicity of the vocabulary used in the construction of their definitions facilitate their use in the extraction of semantic information.

Although several studies in this area, MindNet (Richardson et al. (1998)) can be considered the first lexical ontology automatically created from dictionaries. However, there are more recently works (see for instance O'Hara (2005), Nichols et al. (2005) for English and Gonçalo Oliveira et al. (2008) for Portuguese).

[4]http://www.ldoceonline.com

2.3.2 Learning Ontologies from Corpora

Work-based on electronic dictionaries has had some success. However, the latter are limited in the number of entries. Thus there are some researchers, such as Hearst (1992), Riloff and Shepherd (1997), etc. who began to process corpora text, an endless source of knowledge, to extract semantic information, not just for building lexical ontologies, but also to enrich existing semantic resources (see for example Hearst (1998)).

The process of knowledge extraction in dictionaries is done through an approach essentially linguistic, because of its simplicity of structure and vocabulary used.

However, for unstructured text, the scenario is slightly different and there are some limitations and drawbacks in its use to extract and organise knowledge, for example:

- Difficulty in defining textual patterns able to extract all instances of a particular relationship;
- The possibility of transmit the same idea in different ways, which increases the ambiguity of text;
- There are no boundaries on the vocabulary used, especially in corpora that does not belong to a specific domain;
- Many of the nouns and verbs are modified by adjectives, adverbs or through prepositions;
- The existence of anaphora, where entities previously mentioned in the text, are referred by pronouns.

The work that extracts and organises lexical semantic knowledge from corpora uses methods that can be included in the following categories:

- **Linguistic:** based on the identification of textual patterns and linguistic constructions;
- **Statistical:** based on frequency and co-occurrence of words;
- **Hybrid:** where statistical and linguistic approaches are combined.

Combining linguistic with statistical methods is possibly the best way to deal with the limitations of each one of them. Many of the tasks where lexical semantic information is extracted from text use textual patterns that indicate semantic relations (e.g. hypernymy, meronymy, synonymy, causation, manner, etc.). Hypernymy is probably the most studied relationship, besides Hearst (1992), see for example Caraballo (1999) or Herbelot and Copestake (2006) for English or Freitas (2007), for Portuguese).

Original Pattern	Translation/ Adaptation
NP such as {(NP,)*(or|and) NP}	SUB como {(SUB,)*(ou|e)} SUB SUB tal(is) como {(SUB,)*(ou|e)} SUB
such NP as {(NP,)*(or|and)} NP	tal(is) SUB como {(SUB,)*(ou|e)} SUB
NP {, NP}* {,} or other NP	SUB {, SUB}* {,} ou outro(s) SUB
NP {, NP* {,} and other NP	SUB {, SUB}* {,} e outro(s) SUB
NP {,} including {NP,}*{or|and} NP	SUB {,} incluindo {SUB,}*{ou|e} SUB
NP {,} especially {NP,}*{or|and} NP	SUB {,} especialmente {SUB,}*{ou|e} SUB SUB {,} principalmente {SUB,}*{ou|e} SUB SUB {,} particularmente {SUB,}*{ou|e} SUB SUB {,} em especial {SUB,}*{ou|e} SUB SUB {,} de maneira especial {SUB,}*{ou|e} S SUB {,} sobretudo {SUB,}*{ou|e} SUB

Table 2.5: Hearst patterns and there adaptation to Portuguese language.

Hearst patterns

Hearst (1992) propose a list of six lexical-syntactic patterns to hypernym relations. Table 2.5 shows those patterns, as well as their translation/ adaptation to texts written in Portuguese (Baségio (2007)).

Table 2.5 caption:
SUB: noun
NP: noun phrases

Morin and Jacquemin patterns

Morin and Jacquemin (Morin and Jacquemin (2004)) present patterns that enable the extraction of hypernym relations in French language texts. Table 2.6 presents those patterns, as well as their translation to Portuguese, as proposed by Baségio (2007).

Original Pattern	Translation/ Adaptation
{deux|trois...|2|3|4...} NP1 (LIST2)	{dois|três|...|2|3|4...} SUB1 (LIST_SUB2)
{certain|quelque|de autre...} NP1 (LIST2)	{certos|quaisquer|de outro(s)...} SUB1 (LIST_SUB2)
{deux|trois...|2|3|4...} NP1: LIST2	{dois|três|...|2|3|4...} SUB1 LIST_SUB1
{certain|quelque|de autre...} NP1: LIST2	{certos|quaisquer|de outro(s)...} SUB1: LIST_SUB2
{de autre} NP1 tel que LIST2	{de outro(s)}* SUB1 {tal(is)}* como LIST_SUB2
NP1, particulièrement NP2	SUB1, {particularmente|especialmente} SUB2
{de autre} NP1 comme LIST2	{de outro(s)}* SUB1 como LIST_SUB2
NP1 tel LIST2	SUB1 como LIST_SUB2
NP2 {et|ou} de autre NP1	SUB2 {e|ou} de outro(s) SUB1
NP1 et notamment NP2	SUB1 e (notadamente|em particular) SUB2

Table 2.6: Morin Jacquemin patterns and there adaptation to Portuguese language.

Table 2.6 caption:
SUB1,SUB2: noun
NP1,NP2: noun phrases
LIST_SUB: set of nouns
*: the text before the symbol (*) is not mandatory in the pattern identification

Hearst and Morin-Jacquemin patterns have a vital role in finding hypernym relation between words in text; whether it is domain text, free text or even electronic dictionaries.

However, knowledge extraction works are not limited on hypernymy extraction, but also on other relations, such as meronymy (see Berland and Charniak (1999); Girju et al. (2003) and Girju et al. (2006)), causation (see Girju and Moldovan (2002) and Khoo et al. (2000)) or manner (see (Girju et al. (2003)).

Discovery of New Patterns

One way to automatically discover new patterns is using the following algorithm, proposed by Hearst (1992):

1. Decide the semantic relationship to be searched (e.g. hypernymy, meronymy, synonymy, etc.).

2. Take a list of related term pairs, handcrafted or taken from a knowledge base, for each relationship previously defined (e.g. *cat-animal* (hypernymy) *processor-computer* (meronymy)).

3. Search for text in the corpus where those pairs of terms occur near one another and record the environment.

4. Find similarities between the text saved in the previous step and hypothesise patterns indicating a relationship of interest.

5. Once a new pattern has been positively identified, use it to gather more instances of the target relations and return to the second step.

2.4 Linguistic Resources

In this section, some linguistic resources that are related to this work are introduced. These resources can be used to test developed NLP programs.

The broad-coverage semantic resources are closely related to this work, since they contain semantic knowledge. They can be browsed to make certain decisions related to IE, IR, WSD, QA, among others NLP tasks, or enriched with our extracted knowledge gathered in this thesis.

2.4.1 Corpora

In this subsection, Portuguese textual corpora will be presented. As we can see, some of the corpora described can be browsed through the interface AC/DC (Santos and Sarmento (2003)), supported by Linguateca, a resource center for computational processing of Portuguese language.

AC/DC was initiated in 1999, and created with the purpose of putting all the resources available by the same Web interface[5], and thus facilitate comparison and ease the access to a corpora.

Since 2000, the annotation of these corpus has been automatically made with PALAVRAS (Bick (2000)).

[5]http://linguateca.pt

Bosque

Bosque contains 186,000 words, taken from the CETENFolha and CETEMPúblico corpus. This is the better annotated corpus in Floresta Sintá(c)tica[6], and the most advised for the researches who focus not so much in the quality but more in the accuracy of the results. Besides, Bosque's annotation has been fully reviewed by linguists, which makes it a very reliable corpus.

CETEMPúblico

CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) includes about 2,600 editions of the Portuguese newspaper Público, between 1991 and 1998, totalling approximately 180 million words in European Portuguese. It was created by the project that originated Linguateca, after the signing of a protocol between the Ministry of Science and Technology (MCT) and the Portuguese newspaper Público in April 2000. There are two versions of the corpus: an annotated by the parser PALAVRAS, divided into 196 files, named CETEMPublicoAnot2006.xxx.txt (where xxx ranges [001-196]); another version just contains the text. This corpus can be used for research and technological development.

Colecção CHAVE

Colecção CHAVE was created by 726 editions from the Portuguese newspaper Público and 730 editions of the Brazilian newspaper Folha de São Paulo. This collection is a result of the participation of the Linguateca in the CLEF[7] (Peters et al., 2009) organisation since 2004. In April of 2007, an annotated version (PALAVRAS (Bick, 2000)) of the corpus was created.

COMPARA

COMPARA (Frankenberg-Garcia and Santos (2002)) is a parallel corpus of Portuguese and English. It consists of a database with original texts in these two languages and their translations, linked phrase by phrase. COMPARA is not available for download, however it can be used online. Searching for a word in the Portuguese, the result will be Portuguese sentences containing the searched word as well their English translation.

COMPARA has about 3 million words, being the biggest parallel revised corpus of Portuguese/ English.

Corpus Histórico do Português, Tycho Brahe

Corpus Histórico do Português Tycho Brahe[8] is an electronic morphological and syntactically annotated corpus, composed by Portuguese texts written by authors born between 1380 and 1845. It currently, has 52 texts (2,406,898 words) and it is available for download, just for non commercial purposes. The corpus is developed with the thematic project Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística.

[6]Set of sentences (corpus) analysed (morfo)syntactic, see http://linguateca.pt/Floresta

[7]http://clef-campaign.org

[8]http://tycho.iel.unicamp.br/~tycho

Natura/Minho

Natura/Minho[9] consists of a collection of texts from the regional newspaper Diário do Minho, under the project Natura.

This project is created by several editions of newspaper Diário do Minho; articles that contain only advertising, crosswords, sports and repeated entries were removed.

This resource originated in 2001, was automatically annotated for the first time in 2008 and over time had some versions, being the current the fifth.

WPT05

WPT05[10] is a collection of 10,509,852 documents in Portuguese ($\approx$ 1 million of those are duplicated), obtained by the crawler of the Tumba![11] search engine, produced by the XLDB Node of Linguateca[12].

The corpus includes contents crawled in 2005, selected according to the following criteria: hosted in a *.pt* domain; written in Portuguese, hosted in a *.com*, *.org*, *.net* or *.tv* domain, and referenced by a hyperlink from, at least, one page hosted in a *.pt* domain.

WPT05 collection and related data are available in multiple formats: RDF/XML that includes metadata and text extracted from each URL; ARC format from the Internet Archive, designed for the specific purpose of preserving Web pages as they were crawled. WPT05 succeeds the WPT03 released in 2004, which is a crawl from 2003 distributed since 2004 by Linguateca.

Comparative Analysis

As seen previously, there are several interesting corpora resources that can be used to extract information. In table 2.7 a general analysis of all is made, placing them side by side. Even though it is not plausible to compare them, the main idea is to have a broader view of all the corpus (e.g. text language, type of annotation, size, availability and the type of text used in their construction).

2.4.2 Broad-Coverage Semantic Resources

Broad-coveraqe semantic resources are those intending to cover an entire language, including lexical ontologies. A thesaurus is a more simple lexical ontology that only deals with the synonymy relation, and sometimes antonymy. However, other types of knowledge bases, with slightly different characteristics, are also presented. These deal not only with words and lexical semantic knowledge, but also with common sense.

Thesaurus

Thesaurus associate lists of words with similar meanings within a specific domain of knowledge. Words within the same domain can be synonyms, or antonyms in some cases.

[9] http://natura.di.uminho.pt
[10] http://xldb.di.fc.ul.pt/wiki/WPT_05
[11] http://xldb.fc.ul.pt/wiki/Tumba!
[12] http://xldb.di.fc.ul.pt/wiki/Linguateca-XLDB

Resource	Language	Annotation	Size	Availability	Type
Bosque	Portuguese/ Brazilian	Manual	186 Thousand Words	Academic	Journalistic
CETEMPúblico	Portuguese	Automatic (PALAVRAS)	1.20 GB	Academic	Journalistic
Colecção CHAVE	Portuguese/ Brazilian	Automatic (PALAVRAS)	0.35 GB	Academic	Journalistic
COMPARA	Portuguese/ English	No	3 Millions Words	Not Accessible	Literary
Tycho Brahe	Portuguese	Yes	2 Millions Words	Academic	Historical
Natura/Minho	Portuguese	Yes	-	Not Accessible	Journalistic
WPT05	Portuguese	No	39.6 GB	Academic	Web

Table 2.7: Comparative view in Portuguese corpus.

TeP TeP (Dias-da-Silva and Moraes (2003)) is an electronic thesaurus built manually for the Brazilian variant of Portuguese, that can be seen as a simplified WordNet. Its basic unit is the synset[13], which represents a set of words that have the same meaning, i.e., are linked by a synonymy relationship. Another relationship that can be found in the TeP is antonymy, which represents the opposite relation, linking words with contrary meanings.

In TeP, all synsets have a grammatical category associated. Through the website[14] (Maziero et al. (2008)) it's possible download and browse through 44,678 lexical units, grouped in 19,888 synsets.

OpenThesaurusPT OpenThesaurusPT[15] is a simplified WordNet of Portuguese. This resource is not created by linguists, but by the whole community, respecting the rules required by the responsible ones (the rules are explained in the OpenThesaurusPT FAQ's[16]).

The recourse contains 13,220 words and 4,076 different synsets. The goal of this project is the creations of a thesaurus for the Portuguese language.

Lexical Ontologies

In the last two decades, there have been efforts to create a large database that represents lexical knowledge, where the words and their meanings are represented along with connections held between them. Lexical databases, lexical knowledge bases or lexical ontologies are some of the names given to the resources resulting from these efforts. Some of them are presented in this section.

In general, their construction is aided by structured information from dictionaries, thesaurus or other textual resources like corpora, being achieved by handcrafting or by automatically acquiring information from text. The structure of these knowledge databases usually follow one of the three formalisms, (introduced in section 2.1.3) to represent their knowledge (logical predicates(Smullyan (1995)), directed

[13]Synset: a set of one or more synonyms
[14]http://nilc.icmc.usp.br/tep2/index.htm
[15]http://openthesaurus.caixamagica.pt/index.php
[16]http://openthesaurus.caixamagica.pt/faq.php

graphs or semantic frames(Fillmore (1982)).

Princeton WordNet Princeton Wordnet (Fellbaum (1998)) is a resource that combines traditional lexicographic information with modern computation, in a lexical resource, based on psycholinguistic principles. A wordnet can be seen as a network of lexical structure, that represents mental knowledge into 'words' and their meaning. Besides having been built manually by linguistics, it is probably the most widely used lexical resource in computational linguistics and NLP tasks, and possibly the most complete and most comprehensive that is available on the Web (O'Hara (2005)). Words in Princeton WordNet are clearly divided into nouns, verbs, adjectives, adverbs and functional words. The basic structure in WordNet is the synset, which is a set of synonyms that can be used to represent one concept. Synsets are organised in a network of semantic relations, such as hyponymy and meronymy (between nouns), and troponymy and entailment (between verbs), see examples in table 2.8.

According to statistics provided by WordNet[17], the database consists of 155.287 unique words and 117,659 synsets.

Relation	**Syntactic Category**	**Examples**
Synonymy	N, V, Aj, Av	(pipe, tube) (sad, unhappy) (rapidly, speedily)
Antonymy	Aj, Av, (N,V)	(wet, dry) (powerful, powerless) (rapidly, slowly)
Hyponymy	N	(sugar, maple) (maple, tree) (tree, plant)
Meronymy	N	(brim, hat) (gin, martini) (ship, fleet)
Troponomy	V	(march, walk) (whisper, speak)
Entailment	V	(drive, ride) (divorce, marry)

Table 2.8: Some relations presented in Princeton WordNet.

EuroWordNet EuroWordNet (Vossen and Letteren (1997)) was a project that ended in June 1999. It was the first attempt to join, in a single WordNet, several European languages (English, Spanish, Dutch and Italian). The main idea was that each language had its own wordnet and all wordnets were linked to the English wordnet by equivalence, thus forming a multilingual resource. EuroWordNet was not public, however some of it's samples were available in the past, but they are not anymore.

WordNet.PT WordNet.PT (Marrafa (2002)) and (Marrafa et al. (2006)) is a database of linguistic knowledge for the Portuguese language, organised according to the EuroWordNet. The project started in 1998, after a protocol signed between Instituto Camões and Centro de Línguistica da Universidade de Lisboa. WordNet.PT[18] has a website where it is possible to find information about the project, such as types of relationships (some of them are presented in table 2.9) or publications. However the online search does not work, giving always the following message: *"Sorry, WordNet is down for maintenance. Please come back later"*.

[17]http://wordnet.princeton.edu/wordnet/man2.1/wnstats.7WN.html
[18]http://cvc.instituto-camoes.pt/wordnet/index.htm

Relations general/specific	
'X é hipónimo (é um tipo) de ...'	'X é instanciado como ...'
'X é hipernimo (é supertipo) de ...'	'X é uma instanciação de ...'
Relations part/whole (meronym/holonym)	
'X é merónimo de ...'	'X é merónimo (parte distinta) de ...'
'X é holónimo de ...'	'X é holónimo de ... (parte distinta)'
'X é merónimo (membro) de ...'	'X é merónimo (porção) de ...'
'X é holónimo de ... (membro)'	'X é holónimo de ... (porção)'
'X é merónimo (matéria) de ...'	'X é merónimo (local) de ...'
'X é holónimo de ... (matéria)'	'X é holónimo de ... (local)'
Relations within the structure of the event(correlation)	
'X agente_instrumento ...'	'X instrumento_resultado ...'
'X instrumento_agente ...'	'X resultado_instrumento ...'
'X agente_resultado ...'	'X agente_paciente/ objecto ...'
'X resultado_agente ...'	'X paciente/ objecto_agente ...'
'X paciente/ objecto_instrumento ...'	'X paciente/ génese_resultado ...'
'X instrumento_paciente/ objecto ...'	'X resultado_paciente/ génese ...'

Table 2.9: Some relations present in WordNet.PT.

According to information available in the WordNet.PT Web interface, the project is in development as a EuroWordNet sub project, yet no more significant information is be presented in addition to the last publications of this resource.

WordNet.BR WordNet.BR (Dias-da-Silva et al. (2002); Dias-da-Silva (2006)) is a lexical resource for the Brazilian variant of the Portuguese. The project began to be developed in 2002 and currently only a part of it is available to the public domain (TeP, see 2.4.2), embracing only synonymy and antonymy relations. Plans for the addition of more relations in the future have been reported in Dias-da-Silva (2006), however specification or details are not public.

MindNet The MindNet (Richardson et al. (1998)) is a project developed by Microsoft, built automatically from text and electronic dictionaries. Its origin is the NLP parser used in grammar checking in Microsoft Word 1997. The parser builds syntactic trees and uses pre-defined rules to extract semantic relations. Much of the disambiguation is done through the parser, in the syntactic and morphological level, while the rest of it uses the domain information of the word (information present in electronic dictionaries and the analysis of relations between different words).

Initially, it was based on the extraction of lexical semantic information from electronic dictionaries, more precisely LDOCE[19] and AHD3[20]. Later, encyclopedias and other types of text were also processed. It contains a large set of syntactic and

[19] http://ldoceonline.com

[20] American Heritage Dictionary 3rd Edition

semantic relationships, such as: *Attribute, Cause, Co-Agent, Color, Deep_Object, Deep_Subject, Domain, Equivalent, Goal, Hypernym, Location, Manner, Material, Means, Possessor, Purpose, Size, Source, Subclass, Synonym, Time, Modi, Part* and *User*.

MindNet has a Web interface[21], the MNEX (Vanderwende et al. (2005)). Table 2.10 present part of the MNEX output after searching for the word *"computer"*.

#	**Path**
1	computer $\longleftarrow Tobj \longleftarrow$ *include*
2	computer $\longleftarrow Tsub \longleftarrow$ *use*
3	computer $\longleftarrow Tobj \longleftarrow$ *connect*
4	computer $\longleftarrow Mod \longleftarrow$ *technology*
5	computer $\longleftarrow Mod \longleftarrow$ *software*
6	computer $\longleftarrow Mod \longleftarrow$ *network*
7	computer $\longleftarrow Mod \longleftarrow$ *scientist*
8	computer $\longleftarrow Mod \longleftarrow$ *user*
9	computer $\longleftarrow Mod \longleftarrow$ *graphic*
10	computer $\longleftarrow Mod \longleftarrow$ *screen*

Table 2.10: First ten *Paths* for *'computer'*.

PAPEL PAPEL[22] (Palavras Associadas Porto Editora Linguateca) (Gonçalo Oliveira et al. (2008), Gonçalo Oliveira et al. (2010b)) is a lexical ontology, completely free, based on the automatic extraction of relations between words, that appear in the definitions of a general language dictionary of Portuguese (Dicionário PRO da Língua Portuguesa da Porto Editora (DLP, 2005)).

The first part of its construction was the analysis of the vocabulary used in the dictionary, in order to draw some conclusions about the relationships that would be possible to extract based on textual patterns used in the definitions. Then, based on the conclusions, semantic grammars were created. Those grammars were capable of extracting various types of semantic relationships, and were then used by the syntactic parser PEN (see section 2.6) to process the dictionary and extract the various relationships between the defined words and the set of words that occur in the definitions (Gonçalo Oliveira and Gomes (2008)).

In a more recent phase (Gonçalo Oliveira et al. (2009)), PAPEL 2.0 was evaluated automatically, by comparing relations of synonymy with TeP and the other relations in corpora, after their translation into textual patterns.

According to Gonçalo Oliveira et al. (2010b), PAPEL 2.0 (released on 17/8/2009) contains about 100 thousand lexical items that are distributed according to table 2.11 and about 217 thousand relations distributed according to the table 2.12. It is possible to browse the latest version of PAPEL in the Folheador, available throw its Web interface: `http://sancho.dei.uc.pt/folheador/`.

Others broad-coverage semantic resources

Cyc Cyc project began in 1984 (Lenat (1995)) as an effort to formalise the common sense knowledge into logical structures. In 2003, it had over 1.6 millions of facts

[21] http://stratus.research.microsoft.com/mnex/Main.aspx
[22] http://linguateca.pt/PAPEL

Category	Simple	Multi-word	Total
Adjective	21,000	1	21,001
Adverb	1,390	0	1,390
Verb	10,195	13,866	24,061
Noun	52,599	3,334	55,933

Table 2.11: Distributional items by grammatical categories in PAPEL.

Relation	Quantity
Causation	7,966
Purpose	8,312
Hypernymy	62,591
Place	849
Manner	1,241
Member_of	6,846
Producer	1,292
Property	24,061
Synonymy	79,161
Part_of	6,543

Table 2.12: The relations of PAPEL and their quantities.

(statements) related and more than 118 thousand terms. Currently, it contains about 500 thousand words, including 15 thousand kinds of relationships, and about 5 millions related facts.

To use Cyc on text, it is necessary to map all the text into Cycl language, however the mapping process is very complex because of all the inherent ambiguity in natural language, which must be resolved before the conversion process. Another factor that hinders the implementation of Cyc in a practical task of knowledge extraction is the lack of its full contents to the general public.

Commonsense knowledge spans a huge portion of human experience, however this is typically omitted from social communications, Cyc tries to inferring that knowledge to a formal commonsense knowledge.

ConceptNet ConceptNet (Liu and Singh (2004b)), was built automatically from Open Mind Common Sense (OMCS) (Singh et al. (2002)) and (Liu and Singh (2004a)), in 2000 and released as a project World Wide Web[23], with the main objective of storing text that anyone wants to share about knowledge on common sense (for example, *"The effect of eating food is ..."*; *"A knife is used for ..."*).

With the support of more than 14 thousand contributors who helped in the project development, OMCS has accumulated 1,026,743 common knowledge sentences in English. ConceptNet is an ontology that represents the common sense knowledge, containing 1,6 million edges, connecting more than 300 thousand nodes across 20 semantic relations (see table 2.13) that were developed based on NLP tools, giving rise to a semantic knowledge network (see a snippet of it in figure 2.3).

Berkeley FrameNet Berkeley FrameNet (Baker et al. (1998)) is a project which constitutes a network of *semantic frames* (Fillmore (1982)), where each *frame* has a direct correspondence with a concept, which can describe an object, state or event.

[23]http://openmind.media.mit.edu

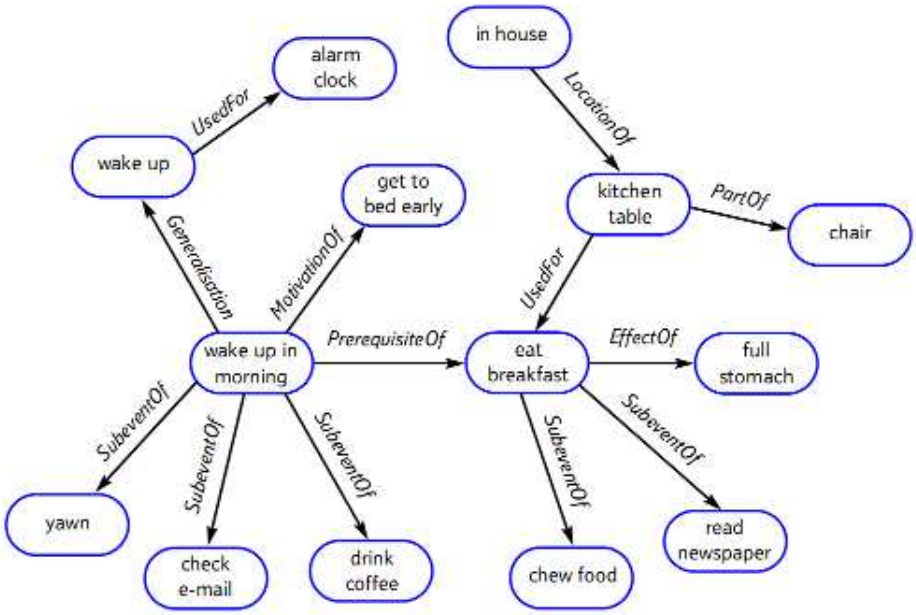


Figure 2.3: An excerpt from ConcetNet's semantic network of commonsense knowledge.

Thematic	Relation-type	Example
K-LINES (1.25 million assertions)	ConceptuallyRelatedTo ThematicKLine SuperThematicKLine	'bad breath' 'mint' 'wedding dress' 'veil' 'western civilisation' 'civilisation'
THINGS (52,000 assertions)	IsA PropertyOf PartOf MadeOf DefinedAs	'horse' 'mammal' 'fire' 'dangerous' 'butterfly' 'wing' 'bacon' 'pig' 'meat' 'flesh of animal'
AGENTS (104,000 assertions)	CapableOf	'dentist' 'pull tooth'
EVENTS (38,000 assertions)	PrerequisiteEventOf FirstSubeventOf SubeventOf LastSubeventOf	'read letter' 'open envelope' 'start fire' 'light match' 'play sport' 'score goal' 'attend classical concert' 'applaud'
SPATIAL (36,000 assertions)	LocationOf	'army' 'in war'
CAUSAL (17,000 assertions)	EffectOf DesirousEffectOf	'view video' 'entertainment' 'sweat' 'take shower'
FUNCTIONAL (115,000 assertions)	UsedFor CapableOfReceivingAction	'fireplace' 'burn wood' 'drink' 'serve'
AFFECTIVE (34,000 assertions)	MotivationOf DesireOf	'play game' 'compete' 'person' 'not be depressed'

Table 2.13: ConceptNet's twenty relation-types illustrated by examples.

It is through syntactic and semantic relationships that it is possible to represent situations involving participants or other conceptual roles that represent the same *frame*. It contains eight different type of relations between frames: *Inheritance, Subframe, Using, Perspective On, Inchoative Of, Causative Of, See Also and Precedes*

Currently, Berkeley FrameNet stores more than 11,600 *lexical units* of English (verbs, adjectives and names), being 6,800 classified as *Finished_Initial*[24] (e.g. "congestion", "transportation", among others) and contains about 960 semantic relations defined between their *frames*, relating over 150 thousand annotated sentences.

The methodology used in the Berkeley FrameNet construction is composed by the following manual steps: (i) initial generation of *frames* and their frame-elements,

[24]Finished_Initial: Regular annotation completed during FrameNet II

(ii) identification of lexical items belonging to each *frame* (iii) manual extraction in the corpus of the example sentences, and (iv) the annotation of example sentences, with the frame-elements. The main disadvantage of this approach is the fact that advance *frames* and elements of each *lexical item* must be specified, to later annotate the example sentences. Steps (i and iv) takes a lot of time consuming and a team of researchers specialised in the area.

Semantic Resources Analysis

As we have seen, there are several types of broad-coverage semantic resources. Each of them has its own structure, representation, license, etc., and each one has its purpose in language processing. At this point, we intend to make a general analysis of all them, placing them side by side (see tables 2.14, 2.15 and 2.16). Even though it is not plausible to compare because some of these resources are significantly different, the main idea is to have a broader view of all of them.

Table 2.14 presents their form of construction (manual, semi-automatic or automatic) and their availability. Table 2.15 includes the basic structure of each resource, its number of nodes (i.e. base structure instances), its number of unique terms (i.e. number of words not repeated that the resource contains), number of edges (instances of relations that link nodes) and the number of relationship types (number of unique relations). Finally, table 2.16 presents the most relevant relationships for this thesis, more precisely synonymy, antonymy, hypernymy, part_of, causation, purpose, location and manner, in order to give a general overview of their importance.

Resource	**Construction**	**Availability**
Cyc	manual	proprietary
ConceptNet	semi-automatic	academic
EuroWordNet	manual	proprietary
FrameNet	manual	academic
MindNet	semi-automatic	proprietary
OpenThesaurusPT	manual	public domain
PAPEL	automatic	public domain
Princeton WordNet	manual	public domain
TeP	manual	public domain
WordNet.BR	manual	TeP
WordNet.PT	manual	proprietary

Table 2.14: Comparative view on lexical databases (construction and availability).

2.5 Similarity Distributional Metrics

Information retrieval (IR) (Singhal (2001)) is the task of locating specific information within a collection of documents (e.g. Web), or other natural language resources (e.g. MRDs), according to some request. Among IR methods, we can find a large number of statistical approaches based on the occurrence of words in documents. Having in mind the distributional hypothesis (Harris (1970)), which assumes that similar words tend to occur in similar contexts, these methods are suitable, for

instance, to find similar documents based on the words they contain or to compute the similarity of words based on their co-occurrence.

Here, we present some distributional metrics that can be found throughout the literature. In their expressions:

- e_i and e_j correspond to entities, which can be words or expressions, as strings of characters that are compared;
- $C = (d_1, d_2, d_3, ..., d_{|C|})$ is a collection of documents used to calculate the metrics;
- $|C|$ correspond to the number of documents contained by the collection C;
- $P(e_i)$ is the number of documents ($d_n \in C$) where e_i occurs;
- and $P(e_1 \cap e_2)$ is the number of documents where e_i and e_j co-occur.

In the previous expressions, $\{C, n, _{|C|}, i, j\} \in \mathbb{N}$.

2.5.1 Corpus Distributional Metrics

Studying the semantic similarity between words has been part of NLP and IR for many years. Semantic similarity measures have a vital importance in various applications in NLP, such as synonym extraction (Lin (1998b)), word sense disambiguation (Resnik (1999)), language modelling (Rosenfeld (1996)), among others.

In this thesis, we explored five metrics (2.2, 2.5, 2.6, 2.7 and 2.9) based on the word's distribution in a corpus.

Cocitation

The measure of Cocitation, in expression 2.1, was first presented in Small (1973) as a similarity measure between scientific papers, after analysing their references. However, it has been applied to other contexts like the similarity between Web pages (Cristo et al. (2003)). In the original expression (2.1), $P(d_i \cap d_j)$ is the number of documents in the collection ($d_n \in C$) referring both documents d_i and d_j and $P(d_i \cup d_j)$ is the number of documents referring at least to one of the documents d_i and d_j.

$$Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)} \tag{2.1}$$

Still, in the scope of this work, we have adapted this expression to measure the similarity between textual entities, which results in expression 2.2, where $P(e_i \cap e_j)$ is the number of documents containing both entities e_i and e_j and $P(e_i \cup e_j)$ is the number of documents containing at least one of the entities. After this adaptions, the measure of Cocitation can be seen as the *Jaccard* coefficient (Jaccard (1901)), used in statistics to compare similarity and diversity in two sets.

$$Cocitation(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \tag{2.2}$$

TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) (Salton and Buckley (1988)), in expression 2.3, is a popular measure in IR which weights (w) the relevance of a term (e_i) in a document (d_j), $w(e_i, d_j)$. Also, in the following expression, $f(e_i, d_j)$ is the frequency, or the number of times e_i occurs in d_j.

$$w(e_i, d_j) = (1 + \log_2 f(e_i, d_j)) * \log_2 \left(\frac{|C|}{P(e_i)}\right) \tag{2.3}$$

LSA

When measuring similarity between two objects, it is common to describe these objects as feature vectors which can be compared. Each entry of these vectors is a numerical aspect describing the object and, in the context of the similarity of documents or words, can be for instance the relevance of a word, or its occurrences in a context. Then, the simplest way to compare the vectors ($\vec{v}$ and $\vec{w}$) is to use the cosine similarity, presented in equation 2.4.

$$\cos(\vec{v}, \vec{w}) = \frac{\vec{v}.\vec{w}}{||\vec{v}||.||\vec{w}||} \tag{2.4}$$

Latent Semantic Analysis (LSA) (Deerwester et al. (1990)) is a measure typically used to rank documents according their relevance to a query. It is based on the cosine similarity, which can be expanded into expression 2.5, to calculate the similarity between entities in the query and the entities in the documents. For the sake of clarity, expression 2.5 considers a query with only two entities, however more entities could be used. Using this measure, higher ranked documents, which have higher cosine values, are those containing entities more similar to the query. In the calculation of LSA, the weight of each entity in a document ($w(e_i, d_k)$ and $w(e_j, d_k)$) can be obtained using TF-IDF, the number of occurrences of e_i in d_k, or other method to compute the relevance of a word in a document.

$$Lsa(e_i, e_j) = \frac{\sum_{k=1}^{|C|} w(e_i, d_k).w(e_j, d_k)}{\sqrt{\sum_{k=1}^{|C|} w^2(e_i, d_k)}\sqrt{\sum_{k=1}^{|C|} w^2(e_j, d_k)}} \tag{2.5}$$

Lin

Lin (Lin (1998a)) presents a theoretical discussion on the definition of similarity. He proposes a measure which does not assume any kind of domain model as long as it has a probabilistic model and is not defined directly by a formula. Still, the measure is derived from a set of assumptions on similarity – the similarity between two objects is the ratio between the information common to both of the objects and the information needed to describe each one of them. Lin shows the generality of its measure when he applies it to domains that go from the similarity between ordinal values (e.g. *good*, *average*, *excellent*), feature vectors, to word similarity, as well as the calculation of semantic similarity in a taxonomy. Expression 2.6 is Lin's measure applied to the similarity of two terms, based on their distribution in

a corpus. There, the information common to both terms is given by the documents where they co-occur and the information needed to describe them is the sum of the documents where each term occurs.

$$Lin(e_i, e_j) = \frac{2 * \log P(e_i \cap e_j)}{\log P(e_i) + \log P(e_j)} \tag{2.6}$$

PMI-IR

The algorithm called PMI-IR (Turney (2001)) uses Pointwise Mutual Information (PMI) to measure the similarity of pairs of words. More precisely, PMI-IR was used to identify (near) synonym words based on their co-ocurrences in the Web, using expression 2.7, or variations of the latter, tuned for a specific search engine.

$$Pmi(e_i, e_j) = \log_2 \left(\frac{P(e_i \cap e_j)}{P(e_i) * P(e_j)} * |C| \right) \tag{2.7}$$

Sigma (σ)

A completely different metric (Kozima and Furugori (1993)), based on the significance of the words in a corpus, was used to measure the similarity between two words. In expression 2.8, which measures the significace of entity e_i, the number of occurrences of e_i in corpus C is given by $O(e_i, C) = \sum_{j=1}^{N} f(e_j, C)$, where $O(e_i, C) \in \mathbb{N}$. Expression 2.9 computes the similarity between entities e_i and e_j.

$$sim(e_i) = \frac{-\log \left(\frac{f(e_i,C)}{O(e_i,C)} \right)}{-\log \left(\frac{1}{O(e_i,C)} \right)} \tag{2.8}$$

$$\sigma(e_1, e_2) = sim(e_1) * sim(e_2) \tag{2.9}$$

2.5.2 Web Distributional Metrics

The most used techniques to measure the semantic similarity use the redundancy and size of a huge corpus, like the World Wide Web, and the results of search engines to measure that similarity. This is usually done by analysing the number of results returned by those search engines in a specific query.

During this thesis, we explored five distinct metrics (2.11, 2.12, 2.13, 2.14 and 2.15). These are used to perform different tasks such as measuring the semantic similarity between words or entities exploiting page counts and text snippets returned by a Web search engine (Bollegala et al. (2007)) and other tasks like word association (Church and Hanks (1989)) or automatic clustering classification (Cilibrasi and Vitanyi (2007)).

WebJaccard

The Jaccard index, also known as the Jaccard similarity coefficient (Jaccard (1901)), is a statistic measure used for comparing the similarity between sample sets. See

equation 2.10, where A and B represent different sets and the result is the sets intersection divided by the size of their union.

$$J(A,B) = \frac{|P(A \cap B)|}{|P(A \cup B)|} \tag{2.10}$$

To compute semantic similarity using page counts, the Jaccard coefficient is modified to the expression 2.11, where $P(e_i \cap e_j)$ is the number of Web documents containing both entities e_i, e_j and $P(e_i) + P(e_j) - P(e_i \cap e_j)$ represents the union of those entities ($P(e_i \cup e_j)$).

$$WebJaccard(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i) + P(e_j) - P(e_i \cap e_j)} \tag{2.11}$$

WebOverlap

WebOverlap is a natural modification to the Overlap (or Simpson) coefficient created by Simpson (1943).The Overlap was devised to minimize the effect of unequal size of two objects being compared (expression 2.12), having in the denominator only the smaller number $\min(P(e_i), P(e_j))$ of occurrences in the sample.

Fallaw (1979) pointed out that if two objects are of approximately equal size, the Jaccard coefficient would be satisfactory, but if there is a great discrepancy in number of occurrences, the larger object would distort the resulting value so that the degree of the relationship between the two objects would be obscured.

$$WebOverlap(e_i, e_j) = \frac{P(e_i \cap e_j)}{\min(P(e_i), P(e_j))} \tag{2.12}$$

WebDice

Bollegala et al. (2007) proposed a page count based on co-occurrence measure, WebDice (expression 2.13), to compute semantic similarity between two given words or named entities, where the notation $P(e_i)$ and $P(e_j)$ denote the page counts for query e_i and e_j respectively in a search engine. In expression 2.13, $P(e_i \cap e_j)$ denotes the page counts for the conjunction query e_i AND e_j.

Given the scale and noise in Web data, it is possible that two words/ entities (e_i AND e_j) may appear on some pages purely accidentally. In order to reduce the adverse effects attributable to random co-occurrences, Bollegala et al. (2007) predefines a threshold c (e.g. c=5).

$$WebDice(e_i, e_j) = \frac{2 * P(e_i \cap e_j)}{P(e_i) + P(e_j)} \tag{2.13}$$

WebPMI

The WebPMI measure is similar to the equation 2.7, however the corresponding $|C|$ variable in expression 2.14 is N, the number of indexed pages reported by the search engine, $N \in \mathbb{N}$.

$$WebPMI(e_i, e_j) = \log_2 \left(\frac{P(e_i \cap e_j)}{P(e_i) * P(e_j)} * N \right) \tag{2.14}$$

WebNWD

The Normalised Web Distance method (NWD), measures two arbitrary objects from the web, in a manner that it is feature free. It is thus being versatile and independent from its domain, genre and language.

The main thrust in Cilibrasi and Vitanyi (2007) is to develop a new theory of semantic distance between a pair of objects, based on (and unavoidably biased by) a background content consisting of a data base of documents. An example of the latter is the set of pages constituting the Internet.

The Normalised Web Distance presented in Cilibrasi and Vitanyi (2007) is an approximation to the Normalised Information Distance (NID) (Bennet et al. (1998)). NWD uses the page counts returned by a search engine in order to calculate the frequencies of these entities. Some formula restrictions are presented in Cilibrasi and Vitanyi (2007).

More recently, the approach in Cilibrasi and Vitanyi (2009), rests on the idea that *information distance* between two objects can be measured by the size of the shortest description that transforms each object into the other one (see equation 2.15).

$$NWD(e_i, e_j) = \frac{\max(\log P(e_i), \log P(e_j)) - \log P(e_i \cap e_j)}{\log N - \min(\log P(e_i), \log P(e_j))} \tag{2.15}$$

In the previews metrics the returned values should be interpreted by different ways:

- WebJaccard, WebOverlap and WebDice return a value in the interval [0-1]. Higher values representing higher similarities.
- WebPMI $\in \mathbb{R}_{\geq 0}$, with higher values represent higher similarities.
- WebNWD $\in \mathbb{R}_{\geq 0}$, with the values closer to 0 representing the higher similarity.

However, some of these properties can change in some rare conditions, i.e. when search engines return more results in $P(e_i \cap e_j)$ than in $P(e_i)$ and/ or $P(e_j)$. In these cases, if the result diverges from its range, we can clip the value to the closer value of the range (e.g. if WebPMI gives a result below 0, it can be simply set to 0).

2.5.3 Other Metrics

In this topic, metrics that will be referred in the following sections, are presented. In the carry out experiments (section 4), equation 2.16 will be used to correlate two different type of data. Precision and recall, equations 2.17 and 2.18 respectively, will be used to analyse the approach evolution in the section 3.2.1.

Next, they are explained in detail.

Correlation Coefficient (ρ)

In order to calculate correlation coefficient between two or more random variables or observed data values, broad class of statistical relationships, equation 2.16 is normally used.

$$\rho(M_i, E_i) = \frac{\sum_i \left(M_i - \overline{M_i}\right)\left(E_i - \overline{E_i}\right)}{\sqrt{\sum_i \left(M_i - \overline{M_i}\right)\left(E_i - \overline{E_i}\right)}} \tag{2.16}$$

More precisely, the correlation coefficient (ρ), will be used to returns the correlation coefficient between two arrays, M_i and E_i, where $\{M_i, E_i\} \in \mathbb{R}$, $\rho \in \mathbb{R}$: $-1 \leq \rho \leq 1$ and $i \in \mathbb{N}$ corresponds to the matrix index.

Precision

In an information retrieval scenario, precision, equation 2.17, evaluates the quality of information extracted. More specifically, is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

$$Precision = \frac{Correct\_answers}{Given\_answers} \tag{2.17}$$

Recall

In an information retrieval scenario, recall, equation 2.18, evaluates the quantity of information extracted. More specifically, recall is defined as the number of relevant documents retrieved by a search, divided by the total number of existing relevant documents, which should have been retrieved.

$$Recall = \frac{Correct\_answers}{Possible\_answers} \tag{2.18}$$

2.5.4 Metrics Applications

Methods for constructing lexical resources only by the identification of textual patterns, similar to Hearst (1992), despite being recurrent, have several problems (Agichtein and Gravano (2000)). Many techniques have been proposed to improve them. For instance, taking advantage of other linguistic constructions (e.g. noun coordination) to improve extraction recall (Roark and Charniak (1998); Cederberg and Widdows (2003)). Others (Agichtein and Gravano (2000); Snow et al. (2005)) propose improving recall and reducing the human effort by using a small set of seed instances or a few handcrafted extraction patterns to make the systems learn extraction patterns. Another alternative is to use a huge corpus such as the Web (see for instance Agichtein and Gravano (2000); Turney (2001); Etzioni et al. (2004) and Cimiano and Wenderoth (2007)) to extract substantially more information.

However, these recall improvement measures tend to reduce the extraction precision. When it comes to improving the latter, distributional metrics are usually a good option to rank the triples based on the similarity between related entities.

While we have calculated all the metrics over a corpus, some of them can be adapted to target the Web and use the hits of a query on a search engine. The PMI-IR is one example of such a metric and some other metrics of this kind (see section 2.5.2) are presented in Oliveira (2009) and in Cimiano and Wenderoth (2007).

As mentioned in the text above, PMI-IR was first developed to identify (near) synonym words. For that specific task, it seemed to perform better than LSA (see Turney (2001)). KnowItAll Etzioni et al. (2004) uses PMI-IR to compute the likelihood that some entity belongs to a class. PMI is calculated as the ratio between the search engine hit counts of each extracted instance and automatically generated indicative textual patterns (e.g. Hearst patterns) associated with the class.

Adaptations of LSA using a *term-term* matrix instead of a *term-document* have also been used to weight relational triples according to the distributional similarity of their arguments (Caraballo (1999); Cederberg and Widdows (2003); Wandmacher et al. (2007)) which can be used to discard triples whose arguments are unlikely to be related (see section 4.1.2).

Lin's similarity measure, adapted to measure the similarity between two synsets, is used in Pantel and Lin (2002) to select the most suitable Wordnet synset for a group of related words extracted from text using a clustering algorithm.

Blohm et al. (Blohm et al. (2007)) study the impact of using several distributional measures (including PMI-IR) to reduce the noise in information extracted from the Web through pattern learning algorithms. In their experiments, an evaluation measure that considers the number of seed pairs of words that produced each learned pattern was the one which performed better. The latter measure favoured more general patterns and penalised patterns which just held for a few examples.

Besides weights assigned according to distributional metrics, the number of times each triple was extracted is usually a good indicator not only of the correction of the triple, but also of its relevance. This hint is also used in several works (Etzioni et al. (2004); Wandmacher et al. (2007)).

2.6 Tools and Libraries

In this section, we present some tools and libraries, that will be studied and analysed in order to understand their possible application in this work.

jSpell

jSpell[25] (Simões and Almeida (2002)) is a morphological analyser that derives from the spell checker ispell. It is implemented in C++. Its main aim is to be used as a spell checker in the Portuguese language. However, besides the information given by the morphological analysis, it can also be used to draw some conclusions about the words analysed (e.g. set of possible interpretations where the lemma can be found, etc.). Dictionaries in jSpell are external to the program. Besides Portuguese, there are dictionaries for other languages (e.g. English). The morphological rules used in the jSpell can be created or changed.

Linguistica

Linguistica[26] (Goldsmith (2001)) is a program (implemented in C++), that can be used to explore unsupervised learning of natural language, focusing primarily on morphology, or word structure. The unsupervised learning in this case refers to

[25] http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell

[26] http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000

the discovery of morphemes of the words: suffix, prefix and lemma. The program is recommended only for Indo-European languages that have at most two suffixes per word. The input should contain a reasonable number of words (minimum 5.000 words), because Linguistica implements statistical methods and heuristics that needs a large quantity of information to begin producing reasonable results.

PTStemmer

PTStemmer[27] is a program that implements two stemming algorithms, Orengo (Orengo and Huyck (2001)) and Porter (Porter (1980)), in order to find the *stem* of the words. PTStemmer is implemented in Java and its purpose is to help information retrieval (IR) systems, i.e., this process can increase IR system coverage (number that measures the quantity of documents that return a query), for example: in a query about *libraries*, the search engine also find documents/ pages with information about *librarians*, because the stem of the word is the same, *library*.

Forma

Forma (Schmid (1994)) is a public domain tool, that make POS tagging of text written in Portuguese.

POS is the most probable morphological category to each word. Lemmatisation provides morphological normalisation. The lemma of a verb is its infinitive form, the lemma of a word (excluding verbs) is itself in the singular form (number), male (gender), when its identification is possible. Forma Web page[28] allows document submission (up to 10Kb) returning the document fully annotated.

FreeLing

FreeLing[29] (Atserias et al. (2006)) is a public domain (GPL licence) linguistic program and can support dictionaries from different languages (Portuguese, English, Spanish and Italian). FreeLing was created in the Universitat Politècnica da Catalunya, in the TALP Research Center and provides morphological and syntactical analyses. There is a Web page here it is possible to test the program `http://garraf.epsevg.upc.es/freeling/demo.php`.

PALAVRAS

PALAVRAS (Bick (2000)) is a morpho-syntactic analyser for Portuguese, that uses information from various linguistic levels: morphological, syntactic and semantic, making a deep document analysis. This information is provided by labels, which provide characteristics of a particular word or sentence structure. The system output provides three different formats: a visual format, a proprietary format parser (VISL)[30] and a TigerXML[31] format. This tool is not available to the public domain

[27] http://code.google.com/p/ptstemmer
[28] http://inf.pucrs.br/ linatural/projetos.html
[29] http://lsi.upc.edu/ nlp/freeling
[30] http://beta.visl.sdu.dk
[31] http://strategoxt.org/Tiger/TigerXML

but it can be used remotely through a Web interface[32] or a text file (up to 2Mb) can be uploaded, being then returned duly annotated.

PEN

PEN[33] (Gonçalo Oliveira and Gomes (2008)) is a generic parser which is implemented in Java according to the description of the Earley's algorithm (Earley (1983)), allowing grammars for different purposes (not necessarily NLP), making it very versatile. The parser input is composed by two text files: one is the input text to be analysed and the other a grammar file. PEN analyses the provided text, line by line and returns the tree(s), regarding the derivation of the line, according to the grammar specification.

PEN is easily integrated into any project and its API allows manipulating and exploiting the derivation trees in order to extract from them the desired information. It was used, for example, in the extraction process that led to the lexical ontology PAPEL (Gonçalo Oliveira et al. (2008)).

QTag

QTag[34] is a POS tagger implemented in Java and is freely available for download (only with resources for English). QTag reads a text and gives each word its corresponding morphology label. It uses a probabilistic method to classify words. This method is independent from the language, and it is desirable that a reasonable number of words (e.g. corpus) is provided to archive a reasonable precision.

Tree-Tagger

Tree-Tagger[35] (available in the public domain) is a morphosyntactic analyser for Linux implemented in Perl.

Pablo Gamallo create additional modules in the Tree-Tagger to supports Portuguese and Galician. The Tree-Tagger analyses the text and it returns their tokens (word), tag (grammatical category of the word) and their lemma.

OpenNLP

OpenNLP[36] is an organisational center for open source projects related to NLP.

It hosts a variety of java-based NLP tools which perform: sentence detection, tokenization, POS, chunking[37] and parsing, named-entity detection, and co-reference using the OpenNLP Maxent [38] machine learning package.

[32] http://beta.visl.sdu.dk/visl/pt/parsing/automatic
[33] http://code.google.com/p/pen
[34] http://phrasys.net/uob/om/software
[35] http://ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[36] http://opennlp.sourceforge.net
[37] In NLP, 'chunking up' refers to moving to more general or abstract pieces of information. While 'chunking down' means moving to more specific or detailed information.
[38] http://maxent.sourceforge.net

OntoLP

OntoLP (Junior (2008)) is a plug-in for Protégé[39] (an open source ontology editor and knowledge-base framework, developed in java), which aims to assist the user during the initial stages in ontology creation, i.e. extraction of candidate terms fro creating concepts and hierarchical organisation of them. For a hierarchical organisation of terms, the plug-in needs to import a corpus for the extraction.

OntoLP builds the ontology from text, through linguistic information that derives from it. To achieve it, information from three levels is used: morphological, syntactic and semantic; however, it is necessary that text is annotated by PALAVRAS (Bick (2000)) and then represented in the XCES[40] format.

Analysis Tools

As mentioned, there are some tools, each one with its purpose. Most of them embrace several NLP levels, but they are not able to be directly compared. In table 2.17 these tools are presented in an overview of all of them (e.g. its availability, the most relevant tasks, etc.).

2.7 Summary

In this section, we have firstly started introducing the basic levels involved in the NLP, and some tasks where they are used, section 2.1.

Related to the scope of this work, in section 2.2, we have introduced some fundamental concepts related to ontologies, such as its definition, categorisation, construction, applications and, at lastly the difference between general ontologies and lexical ontologies.

Since the creation or the augmentation of a lexical ontology is one of our goals, we have studied how the latter are normally created. Usually, they are created with the knowledge extracted from electronic dictionaries or from corpora, section 2.3. However if, on the one hand, there are advantages in using dictionaries, because they are already structured in the format word/ meaning and they use a simple vocabulary, this kind of resource has limited knowledge, is normally static and is not always is available for investigation purposes. On the other hand, at the present it is possible to find a lot of texts in the Web about any possible subject, but their processing is not simple because they have less syntactic restrictions and they use a more varied vocabulary, which can raise more ambiguity problems. There is a third type of resource, which can be classified as semi-structured, that is the encyclopedia. Encyclopedias have also entries for different entities, but their descriptions are larger, and can be considered as text from corpus. Besides that, the encyclopedia's content is not only about words and includes knowledge about the world and human knowledge.

As we will see in section 4.1 and 4.2, a corpus and an encyclopedia, respectively, were used in order to extract semantic knowledge from them, and to create two knowledge sets; but before that, we have analysed same tools and libraries, section 2.6, in order to understand their possible integration in our system.

[39] http://protege.stanford.edu
[40] http://xces.org

Linguistic resources, such as Portuguese corpora and broad-coverage semantic resources were examined in section 2.4. This last one covers thesaurus, lexical ontologies among other semantic resources, all important to make some specific decisions related to IE, IR, WSD, QA, among other NLP tasks.

At last, we have presented some distributional metrics that can be found throughout the literature. Taking in consideration the distributional hypothesis (Harris (1970)), which assumes that similar words tend to occur in similar contexts, we have computed the distributional metrics between words, in order to study their co-occurrence and similarity in text. More specifically, we will use these metrics to validate semantic knowledge. See the next chapter to understand how it was done.

Resource	Core Structure	Nodes	Unique Terms	Relations Instances	Relat. Types
Cyc'03	concepts	118 thousand	-	5 million	-
ConceptNet 2.0	concepts	300 thousand	-	1.6 million	20
FrameNet	frames	11,600	-	150 thousand	-
MindNet'98	words and their dictionary senses	191 thousand	159 thousand	713 thousand	45
OpenThesaurusPT 2.0	synsets	4,076	13,220	-	1
PAPEL 2.0	terms	100 thousand	-	217 thousand	25
Princeton WordNet 3.0	synsets	117,659	155,287	206,941	6
TeP 2.0	synsets	19,884	44,678	4,276(antonyms)	2
WordNet.BR	synsets	44 thousand	18,500	-	2
WordNet.PT v1	synsets	9,015	10,931	11,584	6

Table 2.15: Comparative view on lexical databases (core structure, nodes, unique terms, relation instances and relation types).

Resource	Synonymy	Antonymy	Hypernymy	Part_of	Causation	Purpose	Place	Manner
Cyc'03	yes	yes	yes	yes	yes	yes	yes	yes
ConceptNet 2.0	no	no	yes	yes	yes	yes	yes	no
FrameNet	-	-	yes	yes	yes	yes	-	-
MindNet'98	yes	no	yes	yes	yes	yes	yes	yes
OpenThesaurusPT 2.0	yes(synsets)	no	no	no	no	no	no	no
PAPEL 2.0	yes	no	yes	yes	yes	yes	yes	no
Princeton WordNet 3.0	yes(synsets)	yes	yes	yes	yes(between verbs)	no	no	no
TeP 2.0	yes(synsets)	yes	no	no	no	no	no	no
WordNet.PT v1	yes(synsets)	yes	yes	yes	yes	yes	yes	yes

Table 2.16: Comparative view (existing relations).

Tool	**P. Lang.**	**Language**	**Availability**	**Assignment**
jSpell	C++	Portuguese/ English	Public Domain	Morphological Analysis
Linguistica	C++	Any	Public Domain	Morphological Analysis
PTStemmer	Java	Any	Public Domain	Stemming
Forma	C	Portuguese	Public Domain	POS + Lemmatisation
FreeLing	C++	Portuguese/ Spanish/ Italian	Public Domain	POS
PALAVRAS	Java	Portuguese	Proprietary	POS + (Morpho-syntactic and some Semantic Analysis)
Pen	Java	Any	Public Domain	Syntactic Analysis according to supplied grammars
QTag	Java	Any	Public Domain	POS
Tree-Tagger	Perl	Portuguese/ Galician	Public Domain	POS
OpenNLP	Java	any	Public Domain	POS + Lemmatisation + Tokenization + NER
OntoLP	Java	Portuguese	Public Domain	Semantic Analysis

Table 2.17: Comparative analysis of several tools.

Chapter 3

System Architecture

As referred in the chapter 1, the main goal of our work is to create a system that extracts automatically knowledge from text. After that, another goal is to analyse the benefits of applying metrics to this knowledge, based on the occurrence of words and their neighbourhoods in documents.

To do that, we have created a system with a modular architecture, where each module is independent from each other. Each module performs a specific task on its input text file and outputs another text file, making the maintenance of the system and their debugging much easier. Also, not all modules are mandatory[1].

In this section, an overview of each of these modules is presented. Some of them are tools or libraries integrated in programs created in Java, for a specific task of the system, called as modules. Integrating all of these modules, we obtain a powerful system that extracts and quantifies lexico-semantic knowledge from different sources. All these models are presented in figure 3.1.

The first topic, **Data Extraction**, embraces all necessary tools and libraries to extract text from textual resources, such as *pdfs*, *docs* etc. The **Knowledge Extraction** contains all necessary modules to extract semantic knowledge from text. At last, the **Knowledge Validation** topic is responsible to quantify the knowledge extracted from the previews topic.

The process work-flow is simple, firstly we give some input documents or even a collection of documents to the system, and the **Data Extraction** extracts the text contained in these input files, creating an output file. Then, the **Knowledge Extraction** extracts semantic knowledge, represented as relations triples, from that file, using a parser, grammars among other programs and modules. After that, the **Knowledge Validation** validates the extracted semantic knowledge and weights its triples. Finally, in order to analyse the quantity of new knowledge extracted by our system, the extracted knowledge can be compared with the knowledge in other triple-based knowledge-bases.

The **Data Extraction** topic is explained in section 3.1. In section 3.2, the **Knowledge Extraction** approach is presented, with all underlying modules. Finally, before explaining our idea of how use the extracted knowledge, **Comparing the extracted Knowledge**, section 3.4, we explain how it could be quantified,

[1]For example, in some cases we do not want lemmatise the words, so the Lemmatiser is not needed. If we use text already annotated, the POS tagger would not be necessary. Also, the **Data Extraction** modules can not be necessary if we use a corpus or even a encyclopedia as a textual resource input. Finally, it is not mandatory the inference of hypernymy triples.

Knowledge Validation, section 3.3.

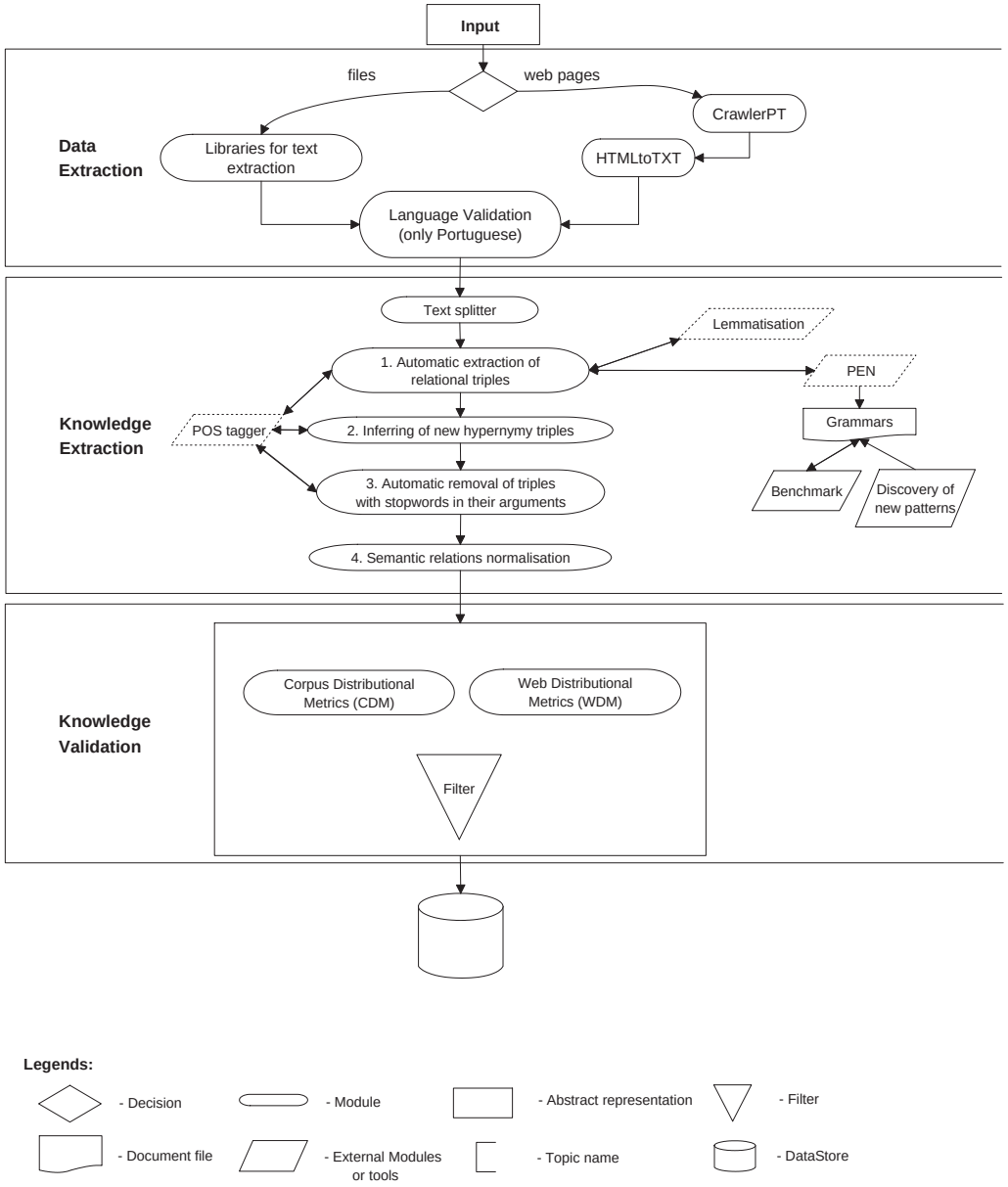


Figure 3.1: System module description.

3.1 Data Extraction

In order to perform the main system task, first it is necessary to build modules to interpret the text contained in different textual sources (e.g. "*.doc", "*.docx", "*.txt", "*.html", "*.rdf", "*.xlsx", etc.). These modules are represented in figure 3.1 as: **Libraries for text extraction**, since they extract text from different file extensions, and **HTMLtoTXT** which, as it name suggests, extracts all the sentences contained in *html* files to *txt* files, excluding HTML tags and other kinds of metadata[2].

[2]Metadata is loosely defined as data about data.

To download data from Web pages for forward use in our system as a corpus, a crawler, named **CrawlerPT** was developed. It starts with a list of URLs, called seeds, and visits each one, identifying all the hyperlinks in those pages and adding them to a new list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. In this case, the most relevant policy is the maximum number of pages that we want crawl.

The next step is to filter only pages or files written in Portuguese, in the module **Language Validation**. This is performed with a simple method: (1) we manually created a list of stopwords[3] for Portuguese, English and Spanish; (2) the program searches in the file for occurrences of those sets of words (if none of them occurs, the file is considered 'unknown'); (3) the file language is identified by the list with more matches (if the Portuguese list has less than 10 matches, the file is not accepted). English and Spanish identification is exploited with the intention of, in future, being used to create a corpus or even extract knowledge from it. The **Data Extraction** topic, represented in the figure 3.1, includes all the modules referred above. As referred above, all the modules contained in the **Data Extraction** can not be necessary, if a corpus or even a encyclopedia is used as a textual resource input.

3.2 Knowledge Extraction

The knowledge extractor phase is a very close adaptation of the current version of a relation extraction system which is currently being developed. More precisely, this system is part of the project Onto.PT[4] (Gonçalo Oliveira and Gomes (2010)), in which development we have also taken part.

Since the system was created to analyse text, each sentence at a time, the preliminary phase of the **Knowledge Extraction** is the **Text splitter** module, that separates text into sentences. It prepares the given text for the extraction phase, where it will be processed according to the grammars, manually created (section 3.2.1) specifically for the extraction of semantic triples between entities[5]. The extracted semantic relations are: synonymy (SINONIMO_DE), hypernymy (HIPERONIMO_DE), part_of (PARTE_DE), causation (CAUSADOR_DE) and purpose (FINALIDADE_DE).

Following, we explain the extraction approach. It follows four stages, which result in a set of relational triples T, that will be used to study the application of the similarity distributional metrics (see section 3.3). These similarity distributional metrics were presented above, in section 2.5.

1. **Automatic extraction of relational triples**: each sentence of a textual input is analysed by the PEN parser, according to the semantic grammars and a triple set, $T = (t_1, t_2, ..., t_n)$, $t_i = (e_1, r, e_2)$ is obtained (see section 3.2.1),

[3]Stopwords are general and very frequent words, usually functional, like prepositions, determiners or pronouns.

[4]http://ontopt.dei.uc.pt

[5]The grammars define each entity either as a simple word, or as words modified by adjectives (e.g. "*Boa* casa" and "Homem *forte*", in English "*Good* house" and "*Strong* man") or by prepositions (e.g. "Garrafa *de* água" and "Raparigas *com* estilo"; in English "Bottle *of* water" and "Girls *with* style").

where $\{_{n,i}\} \in \mathbb{N}$ and $\{s, p, c, h, f\}$[6] $\in r$, where r is the representative relation of interest.

Each entity argument of a triple, e_1 or e_2, can be transformed into is lemma[7].

2. **Inference of new hypernymy triples**: this can be performed by one of the following:

   (a) Inspired on the Costa and Seco (2008) report, all complex entities of the type *noun preposition noun*[8] [**N PREP N**] are analysed. For each entity of this type a new hypernymy triple is extracted, for example if the entity e_1 contained in the triple $t_i = (e_1, r, e_y)$ is a complex entity, the new triple $t_j = (e_x, h, e_1)$ is created, where $\{e_y, e_x\}$ are simple entities and $\{_{i,j}\} \in \mathbb{N}$. Table 3.1, presents same real examples.

Entity	**New triple**
casa_de_campo	*casa* HIPERONIMO_DE *casa_de_campo*
country_house	*house* HYPERNYM_OF *country_house*
garrafa_de_água	*garrafa* HIPERONIMO_DE *garrafa_de_água*
bottle_of_water	*bottle* HYPERNYM_OF *bottle_of_water*

Table 3.1: Examples of triples extracted from multi-word entities.

   (b) Searching for multi-word terms in the text. This method is explained in detail in section 3.2.3.

After performing the inference, the new triples are added to T.

The 2a) method can be considered as our first approach to extract this kind of knowledge. The 2b) method can be considered more comprehensive, because uses all input data, unlike the 2a) that is performed only in the knowledge already extracted.

Nevertheless, these methods have a problem. For example, as Freitas (2007) points out, for the multi-word *pé de atleta*, the new triple *pé* HIPERONIMO_DE *pé de atleta* is wrongly inferred. Were *pé*, in English *foot*, is the lower extremity of the vertebrate leg that is in direct contact with the ground and *pé de atleta*, in English *athlete's foot* is a common fungus infection between the toes in which the skin becomes itchy and sore, cracking and peeling away.

Another problem is that some of the triples obtained by these methods can be too generic or obvious. For example, *sul* HIPERONIMO_DE *sul do Japão* or *república* HIPERONIMO_DE *República Dominicana*, in English *south* HYPERNYM_OF *southern Japan* and *republic* HYPERNYM_OF *Dominican Republic* respectively. This means they may not be relevant to the text meaning.

[6] Initial letters for **S**YNONYM_OF, **P**ART_OF, **C**AUSATION_OF, **H**YPERNYM_OF relation, respectively, and f is the initial letter that corresponds to PURPOSE_OF.

[7] In some cases, multi-word entities that have all its words lemmatised can lead to strange entities.

[8] Only using the preposition *de* or its contraction with an article: *do, da, dos, das.*

However, a method that takes into account the number of occurrences and their utility in collections of documents, is proposed in section 3.3, with the purpose of identifying the triple relevance in the text meaning, using their frequency and their entities co-occurrence. This two modules, 2a) and 2b), are not mandatory, however if they are not used a lot of knowledge would not be extracted/ inferred.

3. **Automatic removal of triples with stopwords in their arguments**: triples with at least one argument in a previously created stopwords list[9], are removed from T, (e.g. the triple *ser vivo* HIPERONIMO_DE *ele*[10] is removed because one of the entities is a stopword, the stopword *ele* in this case).

4. **Semantic relations normalisation**: this stage removes triples if their entities do not respect the category specification (see table 3.2) or change its relation name based on the grammatical category of its arguments. The change in the name of the relation follows a specification, where, for each extracted relation name, there could be a second name according to the grammatical category of its arguments. For example, the triple *carro* SINONIMO_DE *automóvel*[11] will be changed to *carro* SINONIMO_N_DE *automóvel*, because both entities are nouns. Another example: *faltar de profissionalismo* CAUSADOR_DE *grave problema*[12], will be changed to *faltar de profissionalismo* ACCAO_QUE_CAUSA *grave problema*, because in this case a verb *lack of*, causes the effect *serious problem*. See table 3.2 for more details about all category specification. These rules are similar to those applied in PAPEL (Gonçalo Oliveira et al. (2009)), with the intention of a further integration.

Group	**Relation Name**	**Gram. Cat.**
Synonymy	SINONIMO_N_DE	[N,N]
	SINONIMO_V_DE	[V,V]
	SINONIMO_ADJ_DE	[ADJ,ADJ]
	SINONIMO_ADV_DE	[ADV,ADV]
Hypernymy	HIPERONIMO_DE	[N,N]
Part_of	PARTE_DE	[N,N]
	PARTE_DE_ALGO_COM_PROP	[N,ADJ]
	PROP_DE_ALGO_PARTE_DE	[ADJ,N]
Causation	CAUSADOR_DE	[N,N]
	CAUSADOR_DE_ALGO_COM_PROP	[N,ADJ]
	PROP_DE_ALGO_CAUSADOR_DE	[ADJ,N]
	ACCAO_QUE_CAUSA	[V,N]
	CAUSADOR_DA_ACCAO	[N,V]
Purpose	FINALIDADE_DE	[N,N]
	FINALIDADE_DE_ALGO_COM_PROP	[N,ADJ]
	ACCAO_FINALIDADE_DE	[V,N]
	ACCAO_FINALIDADE_DE_ALGO_COM_PROP	[V,ADJ]

Table 3.2: Relations by grammatical category.

[9] The stopwords list was manually created by us to this specific purpose.
[10] *human being* HIPERONIMO_DE *he*
[11] *car* SYNONYM_OF *auto*
[12] *lack of professionalism* CAUSATION_OF *serious problem*

3.2.1 Extracting Knowledge with PEN + Grammars

PEN is a generic parser for which it is possible to create grammars for different purposes (see section 2.6). It allows manipulation and exploitation of derivation trees in order to extract information from text. In our approach, PEN was integrated in the system with that purpose. More specifically, it identifies extract related entities and the name of the relation established between them.

So far, only five different types of semantic relations were exploited: hypernymy, part_of, purpose, causation and synonymy. The six textual patterns proposed by Hearst (1992), for hypernymy relation, were taken in consideration in the construction of the hypernymy grammar, as well as their textual variants. The patterns used by Freitas (2007) and Mineiro et al. (2004), for the extraction of hypernymy from Portuguese text were also considered. For part_of, purpose, causation and synonymy relations, the patterns were created based on:

a) the discovery of new patterns (section 3.2.2);

b) Patterns suggested by Berland and Charniak (1999); Girju et al. (2003), Girju et al. (2006) and Mineiro et al. (2004) for part_of, and Girju and Moldovan (2002) and Khoo et al. (2000) for causation;

c) others pattern variations added by us, after observation.

PEN's input is a string given from the **Extraction triples** module, that needs to be analysed in order to determine if it generates some relevant tree. More specifically, it will be analysed according to every grammar. However, it is possible that PEN does not return any tree, indicating that there are no rules capable to extract any information or the sentence does not have relevant information at all.

On the other hand, one grammar can give rise to more than one derivation for a specific sentence. When this happens and given the Earley algorithm (Earley (1983)), all possible derivations are obtained by PEN. Then, those trees are analysed in the module **Extraction triples**, which automatically decides if they have relevant information or not. If not, they will be ignored.

Let's look at one example.

S(3): *Automóvel é um veículo que é constituído por chassis e motor.*[13]

t_1: *veículo* HIPERONIMO_DE *automóvel*[14]

t_2: *chassis* PARTE_DE *veículo*[15]

t_3: *motor* PARTE_DE *veículo*[16]

For the sentence S(3), two grammars were created, hypernymy and part_of grammars[17], presented in table 3.3 and 3.4 respectively. Based on them, PEN returns several output trees. However, the **Extraction triples** module selects just one

[13] Automobile is a vehicle consisted from chassis and engine.

[14] *vehicle* HIPERNYM_OF *automobile*

[15] *chassis* PART_OF *vehicle*

[16] *engine* PART_OF *vehicle*

[17] Table 3.3 and 3.4 includes just a few rules comparing to system grammars available in `http://ontopt.dei.uc.pt/`

```
RAIZ                    ::= PADRAO <&> QQ

[Regra recursiva, para garantir que a derivação é completa]
QQ                      ::= <?> <&> QQ
QQ                      ::= <?>

PADRAO                  ::= ENUM_HIPONIMOS
                            <&> E_UM_PRES <&>
                            ENUM_HIPERONIMOS

E_UM_PRES               ::= VERBO_SER <&> DET

VERBO_SER               ::= é
VERBO_SER               ::= são

DET                     ::= um
DET                     ::= uma
DET                     ::= uns
DET                     ::= umas

[enumerações]
ENUM_HIPONIMOS          ::= NP_HIPONIMO

NP_HIPONIMO             ::= HIPONIMO

ENUM_HIPERONIMOS        ::= NP_HIPERONIMO

NP_HIPERONIMO           ::= HIPERONIMO_OU_VAZIO

HIPERONIMO_OU_VAZIO     ::= HIPERONIMO
HIPERONIMO_OU_VAZIO     ::= CABECA_VAZIA

CABECA_VAZIA            ::= TIPO
CABECA_VAZIA            ::= PARTE
CABECA_VAZIA            ::= GRUPO
CABECA_VAZIA            ::= NUMERAL
CABECA_VAZIA            ::= EXPRESSAO

HIPONIMO                ::= ENTIDADE
HIPERONIMO              ::= ENTIDADE

ENTIDADE                ::= ENTIDADE_SIMPLES

ENTIDADE_SIMPLES        ::= <?>
```

Table 3.3: Hypernym grammar in the PEN format.

```
RAIZ                    ::= FRAGMENTO
FRAGMENTO               ::= QQ <&> PADRAO <&> QQ

[Regra recursiva, para garantir que a derivação é completa]
QQ                      ::= <?> <&> QQ
QQ                      ::= <?>

PADRAO                  ::= ENUM_HOLONIMOS <&>
                            PADRAO_INVERSO <&>
                            ENUM_MERONIMOS

PADRAO_INVERSO          ::= PADRAO_INCLUI

PADRAO_INCLUI           ::= QUE <&> PADRAO_INCLUI
PADRAO_INCLUI           ::= PADRAO_CONSTITUIDO

QUE                     ::= que

PADRAO_CONSTITUIDO      ::= VERBO_SER <&> CONSTI-
                            TUIDO <&> POR

VERBO_SER               ::= é
VERBO_SER               ::= são

CONSTITUIDO             ::= constituído
CONSTITUIDO             ::= formado

POR                     ::= por

[enumerações]
ENUM_HOLONIMOS          ::= SN_HOLONIMO

SN_HOLONIMO             ::= HOLONIMO

HOLONIMO                ::= ENTIDADE

ENUM_MERONIMOS          ::= SN_MERONIMO <&> E <&>
                            SN_MERONIMO

E                       ::= e

SN_MERONIMO             ::= MERONIMO

MERONIMO                ::= ENTIDADE

ENTIDADE                ::= ENTIDADE_SIMPLES

ENTIDADE_SIMPLES        ::= <?>
```

Table 3.4: Part_of grammar in the PEN format.

derivation tree by grammar, the most complete tree, i.e. the tree with less QQ[18] nodes.

After automatically selecting the most complete tree, the module searches for the hypernym and hyponym nodes, inside the pattern node, named as HIPERONIMO, HIPONIMO and PADRAO respectively, in the hypernym grammar, (see table 3.5). The triple t_1 is then extracted.

In table 3.4 the part_of grammar is presented and the procedure is similar. PEN searches for the holonym and meronym nodes, named as HOLONIMO and MERONIMO in the part_of grammar (see table 3.6), and the triples t_2 and t_3 are extracted.

After all the sentences analysed, a triple set, $T = (t_1, t_2, ..., t_n)$, $t_i = (e_{i1}, r, e_{i2})$ is returned.

The current version of the grammar files are freely available through: `http://ontopt.dei.uc.pt`. Furthermore, details about PEN and grammar creation can be found in Gonçalo Oliveira and Gomes (2008).

To assess the grammars evolution, we soon realised the need of some kind of gold standard, which would be a benchmark standard file, to serve as a basis of comparison for the results that were produced throughout the development.

[18] QQ can be any token, a word, a symbol, etc.

```
[RAIZ]
  [PADRAO]
    [ENUM_HIPONIMOS]
      [NP_HIPONIMO]
        [HIPONIMO]
          [ENTIDADE]
            [ENTIDADE_SIMPLES]
              > [automóvel]
    [E_UM_PRES]
      [VERBO_SER]
        > [é]
      [DET]
        > [um]
    [ENUM_HIPERONIMOS]
      [NP_HIPERONIMO]
        [HIPERONIMO_OU_VAZIO]
          [HIPERONIMO]
            [ENTIDADE]
              [ENTIDADE_SIMPLES]
                > [veículo]
  [QQ]
    > [que]
    [QQ]
      > [é]
      [QQ]
        > [constituído]
        [QQ]
          > [por]
          [QQ]
            > [chassis]
            [QQ]
              > [e]
              [QQ]
                > [motor]
                [QQ]
                  > [.]
```

Table 3.5: PEN output for the sentence S(3) based on hypernymy grammar.

```
[RAIZ]
  [FRAGMENTO]
    [QQ]
      > [automóvel]
      [QQ]
        > [é]
      [PADRAO]
        [ENUM_HOLONIMOS]
          [SN_HOLONIMO]
            [DET]
              > [um]
            [HOLONIMO]
              [ENTIDADE]
                [ENTIDADE_SIMPLES]
                  > [veículo]
        [PADRAO_INVERSO]
          [PADRAO_INCLUI]
            [QUE]
              > [que]
            [PADRAO_INCLUI]
              [PADRAO_CONSTITUIDO]
                [VERBO_SER]
                  > [é]
                [CONSTITUIDO]
                  > [constituído]
                [POR]
                  > [por]
        [ENUM_MERONIMOS]
          [SN_MERONIMO]
            [MERONIMO]
              [ENTIDADE]
                [ENTIDADE_SIMPLES]
                  > [chassis]
        [E]
          > [e]
        [SN_MERONIMO]
          [MERONIMO]
            [ENTIDADE]
              [ENTIDADE_SIMPLES]
                > [motor]
  [QQ]
    > [.]
```

Table 3.6: PEN output for the sentence S(3) based on part_of grammar.

Benchmark

Manual evaluation is the most traditional type of evaluation, however manual evaluation of grammars is a complex task, so a gold standard was manually created to achieve this task.

The golden standard was created manually with some random Portuguese Wikipedia abstracts. These abstracts were manually analysed and a file was created with all plausible triples that could be extracted from those documents[19]. Additionally, with this resource, it is possible to measure the grammar's evolution, however this is only an indicator. To evaluate the results given by the automatic extraction process with the benchmark file, we use two well-know measures, typically used in IR: precision, equation 2.17 and recall, equation 2.18, presented in the section 2.5.3.

3.2.2 Discovery of new Patterns

The discovery of new patterns, presented in figure 3.1 as **Discovery of new patterns**, consists of the identification of new textual patterns that indicate a certain semantic relation. The used approach takes the algorithm proposed by Hearst (see section 2.3.2 *Discovery of New Patterns*) to discover new patterns from text.

After examining several corpus (section 2.4.1), it was decided to use a Portuguese corpus written in European Portuguese, which was large enough to perform the experiment, and created from a context free domain: WPT05 (see section 2.4.1). After choosing the relationships that would be studied and the corpus, rests decide

[19] Anaphora was not take in consideration.

the set of triples that will be used (pairs of words/ entities linked by some semantic relation) in the experiment. PAPEL[20] (see section 2.4.2) has the most suitable option because its triples were extracted with a certain degree of confidence, and are available to the public domain.

In the discovery of new patterns, PAPEL was used, as mentioned above. 750 pairs of entities for each of the three semantic relations hypernymy, part_of, purpose, causation and synonymy, were randomly selected. The search of their co-occurrence only used the first part of WPT05 corpus, due to the fact that is a large corpus 39.6 GB.

In table 3.7, some examples are presented. The number of occurrence is presented in the first column (Occ). In the second column (Extracted Pattern) is presented the textual patterns found between two entities. One random sentence containing the pattern is presented in the third column (Example). Before the last column, where the semantic relation is presented; one (or more) possible inferences of a new indicative textual patterns is suggested. It should also be noted that some of the discovered patterns confirm the Portuguese version of the "Hearst patterns" (Hearst (1992)), which reinforces their importance.

As we can see in table 3.7, some of the discovered patterns have a high occurrence in the corpus, even though they are very ambiguous. For example, the textual pattern $Y|no|X$ does not give relevant information, but together with some verb will indicate the purpose relation.

For instance, consider the verb "usar", in English "to use" with the referred preposition. Besides the well-known ambiguity problem[21], the textual pattern suggested, *(verbo usar) na/no/nas/nos*, can be used to induce purpose, e.g. in the medical domain the medicines are used to something and normally described as X *...used as/for/to...* Y, such as the following example: "anticholinergic *used* orally and locally as an antipruritic"[22].

Another example, the textual pattern $Y|de|X$ can be used to indicate that one entity is used as a purpose of something, but can be used as well as a: connector in Named Entity (e.g. *"Hernani de Jesus da Costa"*), material part (e.g. *"Piso de madeira."*, in English *"Wooden floor."*), (e.g. *"Ele está a norte de Madrid."*, in English *"He is in the north of Madrid."*), among other functions. In the section 3.2.3 we explain how it is possible to extract triples using preposition, applying only some syntactic rules.

Analysing the table 3.7, it is easy to understand the underlying difficulty in the textual patterns creation. Although, conjugating specific verbs with some prepositions, can be seen as a first step in textual patterns creation.

3.2.3 Extracting Triples based on Multi-Word Terms

In this section, a method for extracting semantic relations based on multi-word terms[23] is presented. In our method, if a term occurs modified by an adjective

[20]http://www.linguateca.pt/PAPEL

[21]For example: if we use the pattern *usado no* in the sentence "*Otaku é um termo usado no Japão*" (Otaku is a term used in Japan to describe a fan of a particular subject), the triple *Otaku* PURPOSE_OF *Japão* would be wrongly extracted.

[22]In Portuguese, "anticolinérgicos *usados* por via oral, e localmente como antipruriginoso".

[23]We called multi-word terms, but in the literature sometimes it can be referred as compound nouns. *A compound noun is a noun that is made up of two or more words. Most com-*

Occ.	Extracted Pattern	Example	Patern	Relation
5394	X|do|Y	... *luz* do *sol* ...	da(as/o/os)	Hypernymy
328	X|que são da sua|Y	... *alunos* que são da sua *escola* ...	que é/são da(as/o/os) sua/ seus/ suas/ seus	Part_of
306	X|de|Y	... *camisa* de *dormir* ...	de	Purpose
69	X|e|Y	... *soberania* e *poder* ...	e	Synonymy
61	X|considerei como|Y	... *lugar* considerei como *ponto* ...	(verbo considerar) como	Synonymy
53	X|de|Y	... *acto* de *defesa* ...	de	Hypernymy
44	Y|da nossa|X	... *alunos* da nossa *escola* ...	de(a/os/as) nossa(s)/nosso(s)	Part_of
35	Y|para|X	... *olhos* para *ver* ...	(verbo servir/utilizar/usar) para	Purpose
26	X|sinónimo de noite e|Y	... *jamais* sinónimo de noite e *nunca* ...	sinónimo de(e/os/as)	Synonymy
21	X||Y	... *governo governar* ...	—	Causation
8	X|para|Y	... *voto* para *escolher* ...	(verbo servir/usar) para	Causation
5	Y|no|X	... *livro* no *estudo* ...	(verbo usar) na/no/nas/nos	Purpose

Table 3.7: Discovered patterns from WPT05 corpus.

(e.g. *computador pessoal*, in English *personal computer*) or by a preposition[24] (e.g. *sistema de controlo*, in English *control system*), it will be extracted exactly in that form. This can originate a term with multiple words (e.g. *movimento de massa exclusivo das regiões vulcânicas*, in English *mass movement exclusive from volcanic regions*).

The extraction of hypernymy relations is based on compound terms and takes advantage of two lexical-syntactic patterns, [**N ADJ|ADV**] and [**N de|do|da|com|para N**]. In figure 3.2 the created automaton for this task is presented. For example, from the terms *computador pessoal* (*q0→q1→q3*) and *sistema de controlo* (*q0→q1→q2*) the triples *computador* HIPERONIMO_DE *computador_pessoal* and *sistema* HIPERNIMO_DE *sistema_de_controlo*, are extracted respectively.

In this method, the lexical-syntactic pattern [**N de|do|da|com|para N**] is not applicable if the first **N** is an *empty-head* (Chodorow et al. (1985), Guthrie et al. (1990)). An *empty-head* is a word that does not have any content (e.g. *tipo*, *forma*, in English *type*, *form*) or that implicates the part_of relation (e.g. *parte*, *membro*, *grupo*, *conjunto*, in English *part*, *member*, *group*, *set*). In table 3.3, the set of these words is introduced by the rule HIPERONIMO_OU_VAZIO ::= CABECA_VAZIA, where CABECA_VAZIA, in English *empty-head*. We divided the set of all *empty-head* words in small sets. Table 3.8, presents all these small sets (Group) and some examples (Example).

pound nouns in English are formed by nouns modified by other nouns or adjectives. Cited from: *http://www.learnenglish.de/grammar/nouncompound.htm*

[24]Prepositions are short words (e.g. *de*, *do*, *da*, *dos*, *das*, *com*, *para*, *etc.*, some preposition examples in English: *in*, *on*, *at*, *for*) that usually stand in front of nouns, sometimes also in front of gerund verbs.

Group	Example
PARTE	parte(s) pedaço(s) fragmento(s) membro(s) elemento(s)
GRUPO	grupo(s) conjunto(s) associação(ões) ajuntamento(s) estrutura(s)
TIPO	tipo(s) instância(s) género(s) raça(s) espécie(s)
NUMERAL	dois três quatro
EXPRESSAO	nome(s) termo(s) palavra(s) expressão(ões)

Table 3.8: Some example of words considered empty-heads.

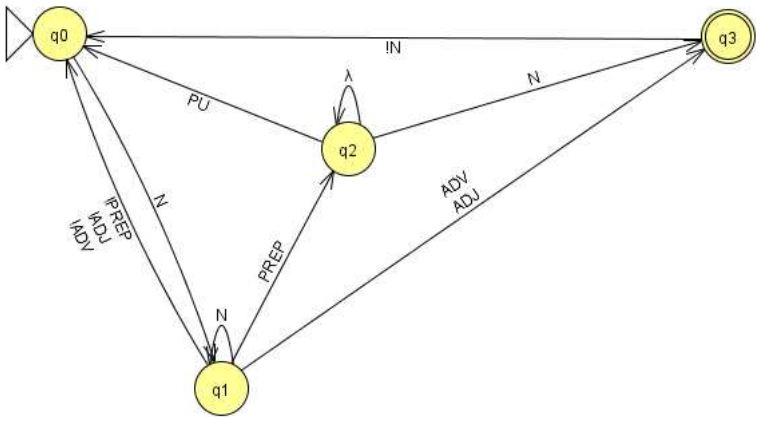


Figure 3.2: Finite-state machine for the extract triples from multi-word terms.

3.3 Knowledge Validation

The task of quantifying knowledge is still an open area and there is not a perfect method. However, there are researchers who apply some specific metrics in particular tasks, such as measuring the similarity between two synsets, identifying (near) synonymous words, etc. (see section 2.5.4).

Nevertheless, we want to create a system that extracts semantic knowledge from a collection of documents or other kind of resource and quantify it automatically. Having in mind the distributional hypothesis (Harris (1970)), which assumes that similar words tend to occur in similar contexts, these methods are suitable for eliminating irrelevant triples or triples whose probability of being correct is low. We think this is a crucial step to have an ontology of high quality, sparing the human effort, typically underlying this kind of tasks.

In order to represent the triple confidence score, we make use of two kinds of metrics: metrics based on words co-occurrence, in corpora (presented in section 2.5.1), and metrics applied to the Web (section 2.5.2). In table 3.9, five abstract

Triple	Weights
e_1 SINONIMO_N_DE e_2 ::	<CDM> ; <WDM>
e_3 PARTE_DE e_4 ::	<CDM> ; <WDM>
e_5 HIPERONIMO_DE e_6 ::	<CDM> ; <WDM>
e_7 CAUSADOR_DE e_8 ::	<CDM> ; <WDM>
e_9 FINALIDADE_DE e_{10} ::	<CDM> ; <WDM>

Table 3.9: Examples of triples internal representation.

triples and its distributional metrics, are presented. There, CDM[25] and WDM[26], correspond to the set of distributional metrics used for corpus and Web respectively. Also, e_i represents an abstract entity, where $i \in \mathbb{N}$: $1 \leq i \leq 10$.

As referred above, the two modules will be used, one for corpus and other for web, represented in the figure 3.1 by the topic **Knowledge Validation**. The module **Corpus Distributional Metrics** gives triples weights (CDM), according not only to their frequency, but also with the output value of the Cocitation, LSA, Lin, Pmi and σ metrics, presented in the equations 2.2, 2.5, 2.6, 2.7 and 2.9, respectively. The second module, **Web Distributional Metrics**, is a module that uses the hits given by a search engine. The metrics used, WebJaccard, WebOverlap, WebDice, WebPmi and WebNWD, are presented in the equations 2.11, 2.12, 2.13, 2.14 and 2.15 respectively (section 2.5.2), and abstractly represented in table 3.9 as WDM.

Applying this methods, we intend to give a confidence value to the set of triples T, automatically.

Furthermore, in order to study how the distributional metrics could be improved, or even create a new set of metrics, each one considering a different type of relation, we intend to use machine learning techniques. With this technique, we intend to filter the set of triples T automatically, with the resulted machine learning thresholds, **Filter** module.

3.4 Comparing the extracted Knowledge

The extracted knowledge, already structured, can be useful to enrich lexical resources, like PAPEL. In that context, it would be interesting to do an analysis on the quantity of extracted knowledge that is not included in the Portuguese resource, just like what Hearst (Hearst (1998)) did for WordNet.

The aforementioned author presents three kinds of outcomes, but just takes the hypernymy relation into account. Still, in our approach we will perform this for all relations of interest. So, to do that, we have changed Hearst's proposal, and we have added new comparison forms.

To understand the next five comparisons, consider DB_a as the knowledge-base that we want to analyse, and DB_b a representative lexical ontology. In this process,

[25] CDM is the set of all corpus distributional metrics values, i.e. $CMD = [CMD_{Cocitation}, CMD_{LSA_o}, CMD_{LSA_t}, CMD_{Pmi}, CMD_{Lin}, CMD_{\sigma}]$, where $CDM_{Cocitation}$, for example, embraces the set of all Cocitation values, that were assigned to T. All CDM values belongs to the $\mathbb{R} \in [0 - 100]$ numbers.

[26] WDM is the set of all corpus distributional metrics values, i.e. $WMD = [WMD_{WebJaccard}, WMD_{WebOverlap}, WMD_{WebDice}, WMD_{WebPmi}, CMD_{WebNWD}]$, where $WDM_{WebJaccard}$, for example, embraces the set of all WebJaccard values, that were assigned to T. All WDM values belongs to the $\mathbb{R}_{\geq 0}$ numbers.

we want to know the number of triples, $t_i = (e_1, r, e_2)$ in DB_a:

C1: already in the database DB_b;

C2: present in DB_b - but with a relation different from r;

C3: that have both entities, e_1 and e_2 in DB_b - however not related;

C4: that just have one of its entities in DB_b.

C5: that do not have any of its entities in DB_b.

So that all events are considered to be mutually independent[27], i.e. $C1 \cap C2 \cap C3 \cap C4 \cap C5 = \varnothing$, C1 and C2 need to be firstly executed. To a better understanding, we propose the algorithm 1, that describes a finite sequence of instructions that need to be followed to perform a non-ambiguous comparison.

Algorithm 1 Comparing common knowledge between to knowledge-bases (DB_a and DB_b).

```
C1, C2, C3, C4, C5 ← 0;
for i = 0 to DB_a.length do
    t_a ← DB_a[i]; //t_a = e1_a, r, e2_a
    e1_a ← t_a[0];
    e2_a ← t_a[2];
    for j = 0 to DB_b.length do
        t_b ← DB_b[j]; //t_b = e1_b, r, e2_b
        e1_b ← t_b[0];
        e2_b ← t_b[2];
        if (t_a = t_b) then
            C1 ← C1 + 1;
            break;
        end if
        if ((e1_a + e2_a) = (e1_b + e2_b)) then
            C2 ← C2 + 1;
            break;
        end if
    end for
    if ((e1_a && e2_a) exist in DB_b) then
        C3 ← C3 + 1;
        break;
    end if
    if (e1_a exists in DB_b) || (e1_b exists in DB_b) then
        C4 ← C4 + 1;
        break;
    end if
    if ((e1_a && e2_a) do not exist in DB_b) then
        C5 ← C5 + 1;
        break;
    end if
end for
```

Nevertheless, this approach entails a problem: if the lexical ontology is based on synsets, it will be not easy to compare the quantity of entities present in both resources. However, the entities in the same synsets, are associated by synonymy relation, and between a synsets by other type of relation of interest.

[27]Two events A and B are mutually independent, if and only, if $A \cap B = \varnothing$, and $P(A \cap B) = 0$.

Chapter 4

Experimental Work

This chapter describes the results of four experiments carried out. The first is an experimental approach in a newspaper corpus (CETEMPúblico), section 4.1. We have used a simple version of our system, and we have studied the possibility of using distributional metrics to improve the precision of relational triples, automatically extracted from Portuguese text.

The second experiment is an approach in a collection of unstructured text (Wikipedia abstracts), using the current version of the system, section 4.2. We improved the **Knowledge Extraction** modules, presented in section 3.2, to perform the automatic extraction of relational triples from the Wikipedia abstracts. Also, some methods were used in order to, in an automatic way, validate and evaluate the knowledge extracted.

In both experiments, the extraction system used are two versions of the Onto.PT[1] (Gonçalo Oliveira and Gomes (2010)). There have been improvements in the second, at three levels: (1) more grammar patterns; (2) POS tagger integration to identify adjectives and to enable the next point; (3) extraction of hypernymy relations from compound nouns. In this versions were used two methods to extract hypernymy relations from compound nouns. The first version (section 3.2), not had the capability of take advantage of compound names, so we have implemented a method to take advantage of the annotated version of the CETEMPúblico corpus (see *Experiment 1*, section 4.1 for more details). In the last version has been used a method that uses the POS tagger that is integrated in the Onto.PT (see *Experiment 2*, section 4.2 for more details).

The third experiment describes a second approach in the CETEMPúblico corpus, comparing the first and the second version of our system, section 4.3.

The fourth experiment, and the last one, presents an experimental approach that analyses the quantity of common knowledge between the three resources, section 4.4.

4.1 Experiment 1: knowledge extraction from CETEMPúblico

This section presents the experiment carried out to study the possibility of using distributional metrics to improve the precision of relational triples, automatically extracted from Portuguese corpora.

[1] http://ontopt.dei.uc.pt

4.1.1 Experiment Goals

The most common ways to evaluate new knowledge involve either manual inspection by human judges or the comparison with a gold standard. However automatic methods have been developed to validate it, see section 2.5.4, based on existing resources (e.g. corpus, Web, etc.), which are usually exploited together with pre-conceived assumptions (e.g. related words tend to co-occur, some relation can be denoted by a set of discriminating textual patterns) and some mathematical formulas to quantify the quality of the new knowledge (section 2.5).

Additionally, in order to calculate the confidence on their results or to improve the precision of knowledge extraction systems, several authors, see section 2.5.4, have taken advantage of distributional metrics, presented in section 2.5.

Having this in mind, the goal of this experiment is to study how existing distributional metrics may be used to improve the quality of information extracted, automatically from text. Additionally, will be studied how these evaluation may benefit from using these metrics.

To study the impact of the latter, we have integrated several metrics, described in section 2.5.1, in our system, that was presented in the previews chapter (3).

The system is based on a set of semantic grammars which include textual patterns that frequently denote semantic relations. Since it has the first experiment, the grammars still very simple, however they were improved in the more recent experiments, more precisely in the *Experiment 2*, section 4.2.

We are aware that it captures an excess of extraneous and incorrect information, especially from unstructured text. However, regarding the goal of this experiment, this is not a problem but an added value, since we explicitly aim to test whether the metrics applied are capable of identifying these situations. Furthermore, using machine learning techniques, we will ascertain if it is possible to come up with a new metric based on one or several existing metrics.

Lastly, Web metrics, presented in section 2.5.3, will be used as a first attempt of study their application in the knowledge validation task.

4.1.2 Experiment

Experiment Set-up

Through this experiment we have used the part-of-speech annotated version of the CETEMPúblico corpus (Santos and Rocha (2001)), provided by Linguateca[2] (see section 2.4.1).

Due to the limitations in the processing time and storage, we ended up using only the first 28,000 documents of CETEMPúblico, which contain 30,100 unique content words (considering only nouns, verbs and adjectives) and results in approximately 1 million of *word-in-document* relations, called *term-document* matrix.

To understand the concept *word-in-document* or simply *term-document* matrix, let's see an example. Consider two small documents, d_1, that contains the first two sentences, S(4) and S(5), and d_2, containing the sentence S(6). It's words are presented in their lemmatised form, and the *italic terms*, represent the used terms, nouns, verbs or adjectives in the *term-document* matrix construction.

[2]http://www.linguateca.pt

The collected document vectors into a *term-document* matrix, is presented in table 4.1.

S(4): *Inez Teixeira ser* um *jovem pintora* que *ter expor* regularmente desde há um dois *ano*.

S(5): O *governo cambojano assinar* um *acordo* de *paz* com três *facção* da *guerrilha*, *incluir* o *khmer vermelhos*.

S(6): No entanto, para já, o *governo português* apenas *ter conhecimento* de um comunicação que *ter ser fazer* por *Jonas Savimbi* ao *Zambia*, a qual *anunciar aceitar* que o *governador* do *Huambo ser nomear* pelo *governo* de *Angola*.

term/document	d_1	d_2
inez	1	0
teixeira	1	0
ser	1	1
jovem	1	0
pintora	1	0
expor	1	0
ano	1	0
governo	1	2
cambojano	1	0
assinar	1	0
acordo	1	0
paz	1	0
fracção	1	0
guerrilha	1	0
incluir	1	0
khmer	1	0
vermelhos	1	0
português	0	1
ter	0	2
conhecimento	0	1
fazer	0	1
jonas	0	1
savimbi	0	1
zambia	0	1
anunciar	0	1
aceitar	0	1
governador	0	1
huambo	0	1
angola	0	1

Table 4.1: *Term-document* matrix example.

A relational database, which can be seen as an occurrence matrix, table 4.1, was used to save this information and also the TF-IDF (weight of the word in the document) of all words.

This occurrence matrix provides:

i) the number of documents, d_k;

ii) the number of times the word l_i occurs;

iii) the documents where l_i occurs;

iv) the number of words in d_k, N_{d_k};

v) the total number of words in the corpus, N;

vi) the relevance R_{l_i} of the word l_i in the corpus.

Where, $\{d_k, l_i, N_{d_k}, N\} \in \mathbb{N}$ and $R_{l_i} \in \mathbb{R}_{>0}$. With this information we can calculate the co-occurrence between l_1 and l_2 and the number of times both occurs $P(l_1, l_2)$, where $P(l_1, l_2) \in \mathbb{N}^0$.

In this experiment, three of four stages described in section 3.2, and presented as **Knowledge Extraction** topic in the figure 3.1, are used, as shows figure 4.1. Being CETEMPúblico a corpus, with structured information, this experiment do not take advantage of the **Data Extraction** modules. **POS tagger** and **Lemmatisation** modules were not necessaries, because the corpus is already annotated.

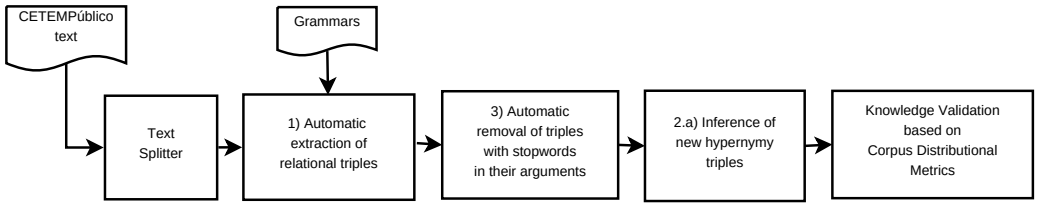


Figure 4.1: Modules used in *Experiment 1.*

Extraction Results

For experiment purposes, extraction was also performed over the first 50,000 documents of CETEMPúblico and a total amount of 20,308 triples was first obtained by the **Automatic extraction of relational triples** module. Then, after the discarding phase, 5,844 triples (28.8%) were removed from the later set by the **Automatic removal of triples with stopwords in their arguments** module. Finally, inference resulted in more 2,492 (17.2%) new triples.

The **Automatic removal of triples with stopwords in their arguments** module, appears before the **2(a) Inference of new hypernymy triples** module in this experiment, because the last one only use complex entities, in the type [**N PREP N**] to infer new triples, and in this case, it is necessary that wrong triples be previously eliminated, to do not propagate possible errors. The 2(a) method is described in section 3.2.

The final triple set included 16,956 triples, more precisely 270 synonymy triples, 9,365 hypernymy, 1,373 part_of, 2,660 causation, and 3,288 purpose. Two example sentences and the triples extracted from them, as well as their translation, are shown in table 4.2. In the second example, one of the problems of the extraction system is in evidence: the parser can only connect the word *diplomat* with *Egypt* and not with the other countries in the enumeration, but an erroneous triple is extracted anyway.

Application of the Metrics

The distributional metrics, contained in the topic **Knowledge Validation based on Corpus Distributional Metrics** in figure 4.1, as previously referred in section 2.5.1, more precisely in expressions 2.2, 2.5, 2.6, 2.7 and 2.9, were implemented and normalised to fit the interval [0-100]. For instance, PMI-IR was normalised based on Bouma's (Bouma (2009)) proposal. Also, calculation of the weights $w(e_i, d_k)$ in the LSA expression (2.5) was done by two different methods: the number of occurrences of entity e_i in the document d_k (LSA_o) and TF-IDF (LSA_t).

Each distributional metric was applied to the triple set, T, in the following manner: for each triple $t_i = (e_1, r, e_2), t_i \in T$ and $_i \in \mathbb{N}$, the distributional similarity between e_1 and e_2 was computed. For multiword entities, the metrics were applied

Sentence	**Triple(s) extracted**
... possibilidade de transplantar para o homem pulmões, rins ***ou outros*** *órgãos colhidos em porcos ...*	(*órgão* HIPERONIMO_DE *pulmão*) (*órgão* HIPERONIMO_DE *rim*)
... the possibility of transplanting to humans lungs, kidneys ***and other*** *organs obtained from pigs ...*	(*organ* HYPERNYM_OF *lung*) (*organ* HYPERNYM_OF *kidney*)
A delegação ***inclui*** *diplomatas do Egipto, Irão, Paquistão, Saudita e Senegal.*	(*diplomata_do_Egipto* PARTE_DE *delegaçcão*) (*Irão* PARTE_DE *delegaçcão*) (*Saudita* PARTE_DE *delegaçcão*) (*Senegal* PARTE_DE *delegaçcão*)
The delegation includes diplomats from Egypt, Iran, Pakistan, Saudi Arabia and Senegal.	(*diplomat_from_Egypt* PART_OF *delegation*) (*Iran* PART_OF *delegation*) (*SaudiArabia* PART_OF *delegation*) (*Senegal* PART_OF *delegation*)

Table 4.2: Extraction examples of triples extracted from CETEMPúblico.

between each word of one entity and each word of the other, in order to calculate the average similarity value.

Manual Evaluation

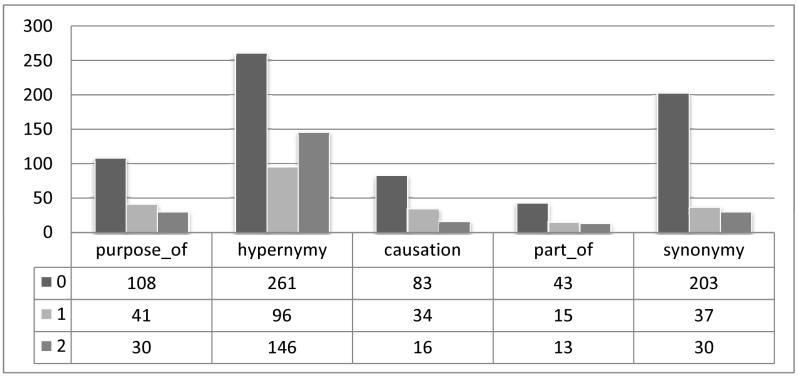
To evaluate the precision of the results, we selected random samples for each type of relation. The samples' sizes took the type of relation into consideration and were the following: 503 hypernymy triples (5.4%), 179 purpose relations (5.4%), 133 causation relations (5.0%), 71 part_of relations (5.2%) and of 270 synonymy relations (100%), totalling 1,156 triples, which were divided into ten random samples, each one evaluated by one of ten human judges.

Each human judge was asked to assign one of the following values to each triple, according to its quality:

- **0**, if the triple is completely incorrect.
- **1**, if the triple is not incorrect, but something is missing. Like a preposition or an adjective that makes one of the arguments strange and prevents the triple from being correct in one or both of its arguments, or even the relation is very generic.
- **2**, if the triple is correct.

A sentence describing the meaning of each relation was provided together with the triples to validate.

The number of triples obtained by manual evaluation are presented in the figure 4.2. As we can see in the y-axis, there are many incorrect triples, which show that the extraction system is far from perfect. Nevertheless, we were expecting to reduce the number of incorrect triples after applying a filter based on one or several distributional metrics.



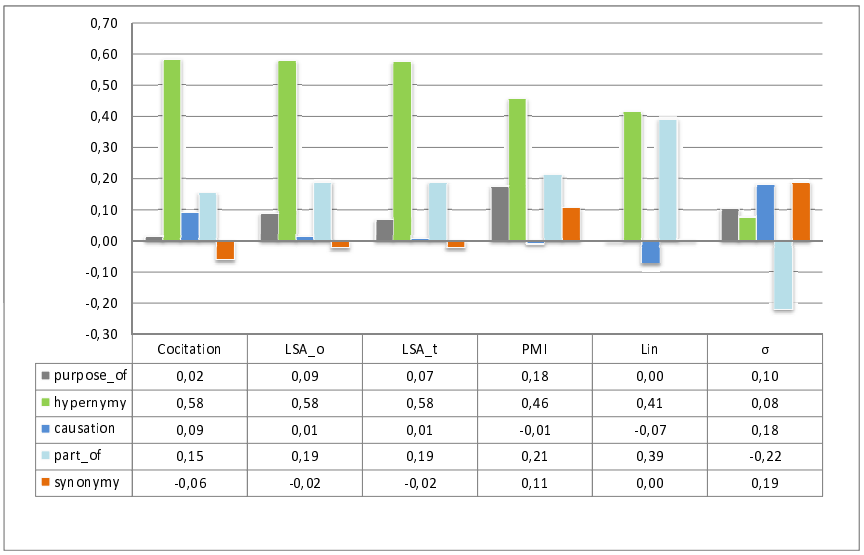
	purpose_of	hypernymy	causation	part_of	synonymy
0	108	261	83	43	203
1	41	96	34	15	37
2	30	146	16	13	30

Figure 4.2: Manual Results.

Manual Evaluation vs. Distributional Metrics

Table 4.4 shows some examples of extracted triples and puts side-by-side their manual evaluation and the calculated metrics. Since the triples were extracted from the same corpus used to obtain the metrics, the latter values are never zero except for Lin's measure in the triple *palavra* HIPERONIMO_DE *beato*. However, this happens because these words only co-occur once and, even though *palavra* is a very frequent word, *beato* is very infrequent.

In order to observe the relationships between the manual evaluation and the output values given by the metrics, the correlation coefficients between them were computed and are shown in figure 4.3.



	Cocitation	LSA_o	LSA_t	PMI	Lin	σ
purpose_of	0,02	0,09	0,07	0,18	0,00	0,10
hypernymy	0,58	0,58	0,58	0,46	0,41	0,08
causation	0,09	0,01	0,01	-0,01	-0,07	0,18
part_of	0,15	0,19	0,19	0,21	0,39	-0,22
synonymy	-0,06	-0,02	-0,02	0,11	0,00	0,19

Figure 4.3: Correlation coefficients between manual evaluation and the distributional metrics.

For this purpose, the correlation coefficient presented in equation 2.16, section 2.5.3, will be used.

The correlation coefficient, ρ, returns the correlation coefficient between two arrays, in this case, the manual evaluation matrix (M)[3] and the automatic values

[3] M contains a set of triples and their manual evaluation.

given by the metrics (E)[4].

It is possible to observe that most metrics are strongly correlated with the quality of the triples, except for synonymy. This happens because all metrics except σ are based on co-occurrences and, in corpora text, synonymy entities, despite sharing very similar neighbourhoods, may not co-occur frequently in the same sentence (Dorow (2006)) or even in the same document because they are alternative names for the same thing. This might also be the reason for the low correlation coefficients with σ, which is based on the relevance of the terms.

Higher correlation coefficients are obtained for the hypernymy relation with the metrics of PMI and, especially, LSA and Cocitation/ Jaccard, which suggests that hyponyms and their hypernyms tend to co-occur more frequently than causes or purposes. Also, there are more ways to denote the latter relations in corpora text which led to less extracted and more incorrect triples. This is in conformity with an experience (Gonçalo Oliveira et al. (2009)) where patterns denoting these relations were looked for in CETEMPúblico to validate semantic triples included in the lexical resource PAPEL. On the other hand, part_of relations have good correlation coefficients with Lin's measure and LSA.

Another conclusion is that, with this experience, the obtained values for LSA calculated with the occurrences of the entities (LSA_o) are very similar to the ones calculated with the TF-IDF (LSA_t). However, calculating the number of occurrences of a term in a document is much faster than computing the TF-IDF.

An additional experiment, taking only triples with simple-entities[5] in their entities vs. the distributional metrics, has performed as shows figure 4.4.

Higher correlation coefficients with all metrics are obtained for the part_of relation, less for σ, because it is not based on co-occurrences. The correlation values suggests that, part_of entities tend to co-occur more frequently than the others, like we conclude in the preview experiment. Nevertheless, the hypernymy correlation values decreases, see figure 4.4, however this can be explained with the fact that their entities are normally constituted by multi-words, and the hypernymy triples with simple-entities tend to co-occur sparsely.

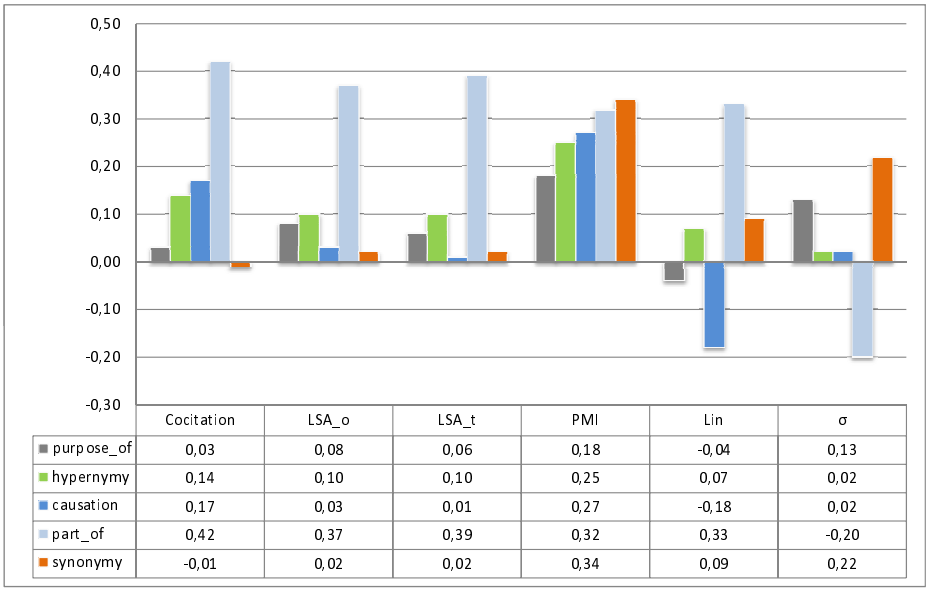
Again, with this experience, the LSA values calculated with the occurrences of the entities (LSA_o), are very similar to the ones calculated with the TF-IDF (LSA_t).

Furthermore, in order to study the possible combination of the distributional metrics to create a new set of metrics, each one considering a different type of relation, we have used machine learning techniques, more specifically the toolkit Weka (Witten and Frank (1999)). Several datasets were created, each one for a different relation. These datasets comprise a set of triple evaluation scores and their manual evaluation, as the entries of table 4.4, and were used for training several classification algorithms.

The best learned modules using the algorithms of isotonic regression and also simple linear regression are shown in table 4.3 together with their correlation coefficient. There are two situations where the modules are not present because the

[4] E is array of all values given by one specific metrics, for example $E = Cocitation = [CDM_{Cocitation1}, CDM_{Cocitation2}, ..., CDM_{Cocitationk}]$, were $CDM_{Cocitationk}$ is the Cocitation value that correspond to the triple t_i, were $\{_{k,i}\} \in \mathbb{N}$.

[5] We call triples with simple-entities to a triple that contains only one word for entity, for example *wheel* PART_OF *car*.



	Cocitation	LSA_o	LSA_t	PMI	Lin	σ
purpose_of	0,03	0,08	0,06	0,18	-0,04	0,13
hypernymy	0,14	0,10	0,10	0,25	0,07	0,02
causation	0,17	0,03	0,01	0,27	-0,18	0,02
part_of	0,42	0,37	0,39	0,32	0,33	-0,20
synonymy	-0,01	0,02	0,02	0,34	0,09	0,22

Figure 4.4: Correlation coefficient between manual evaluation (simple entities) and distributional metrics.

obtained correlation coefficients were very low and it did not make sense to choose the best. As one can see, most of the best results for numeric classification were obtained with the isotonic regression module which picks the attribute that results in the lowest squared error, and defines several cut points, assuming a monotonic function.

Relation	**SimpleLinear**	**Correl**	**Isotonic**	**Correl**
Causation	(0.01*σ+0.05)	0.12	-	-
Purpose	(0.02*PMI-0.6)	0.22	PMI	0.24
Hypernymy	(0.02*Cocitation+0.49)	0.56	Cocitation	0.66
Part_of	(0.01*Lin+0.26)	0.28	Cocitation	0.38
Synonymy	-	-	σ	0.22

Table 4.3: Learned metrics with higher correlation coefficient.

The J48 was the best algorithm for discrete classification. J48 is an improved version of the C4.5 algorithm (Quinlan (1993)) and its result module is a decision tree, such as the one in figure 4.5, obtained using a 10-fold cross-validation test and which classifies 59.1% of the purpose_of triples correctly. As one can see, this tree classifies the triples into one of the following classes, corresponding to the manual evaluation scores (0, 1 and 2).

Evaluation of the Precision

Based on the experiment presented in the previous sections, we have analysed the impact of using a filter based on the best metrics obtained with the isotonic regression (Barlow et al. (1972)). Figures 4.6 and 4.7 present the evolution of the precision using different cut points on the Cocitation/ Jaccard metric for the hypernymy and part_of triples, respectively. In the figures, the x-axis correspond to the threshold

Triple	Manual	Coc	LSA_o	LSA_t	PMI	Lin	σ
livro HIPERONIMO_DE *livro_de_reclamações* *book* HYPERNYM_OF *complaints_book*	2	100	100	100	100	94.85	27.5
nação SINONIMO_DE *povo* *nation* SYNONYM_OF *people*	2	4.21	7.92	8.21	66.65	55.12	35.79
violência CAUSADOR_DE *estrago* *violence* CAUSATION_OF *damage*	2	1.60	4.38	4.47	63.90	29.51	43.82
palavra HIPERONIMO_DE *beato* *word* HYPERNYM_OF *pietist*	1	0.16	1.75	1.78	61.83	0	48.25
poder CAUSADOR_DE *algum_desigualdade* *power* CAUSATION_OF *some_difference*	1	0.27	3.07	3.25	54.82	45.52	26.15
jogo FINALIDADE_DE *preparar* *game* PURPOSE_OF *prepare*	1	1.61	3.53	3.62	50.89	48.22	25.52
sofrer SINONIMO_DE *praticar* *suffer* SYNONYM_OF *practice*	0	0.73	1.34	1.37	52.04	27.77	34.25
atender FINALIDADE_DE *moderno* *answer* PURPOSE_OF *modern*	0	0.69	1.81	1.82	55.22	13.84	41.24

Table 4.4: Examples of extracted triples, their manual evaluation score and their computed distributional metrics.

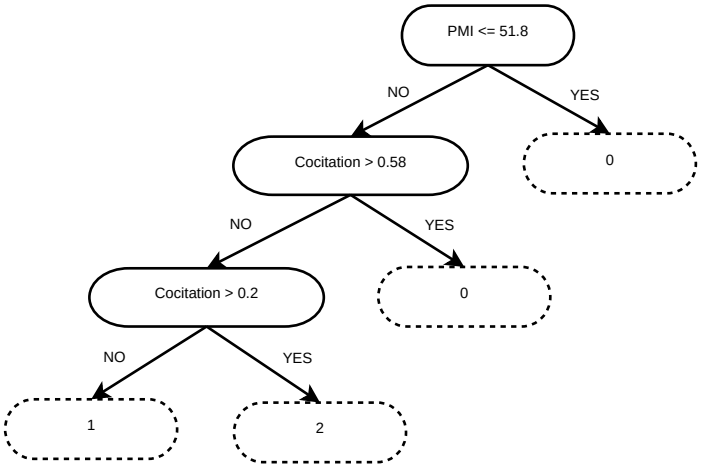


Figure 4.5: The J48 decision tree learned for purpose.

values and the y-axis to the precision values (see equation 2.17 on section 2.5.3 for more information about the precision formula).

Of course that, while the cut point increases, less triples are obtained, but the majority of the discarded ones are wrong, leading to a higher precision. From a certain cut point, the amount of triples starts to decrease giving rise to more variations in the precision. Therefore, after observing the figures 4.6 and 4.7 we would define 50 and 1 as adequate cut points for hypernymy and part_of, respectively. Applying these thresholds, the resulted number of triples from the total evaluated for hypernymy and part_of relations are: 98 (10 evaluated with 0, 10 with 1 and 78 evaluated with 2) and 27 (9 evaluated with 0, 6 with 1 and 12 evaluated with 2) respectively.

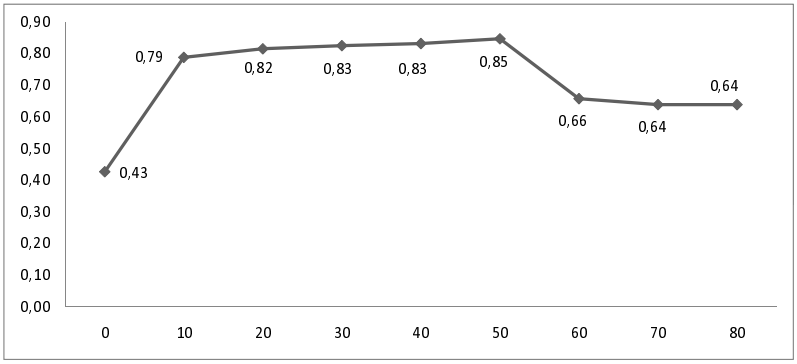


Figure 4.6: Evolution of the precision when increasing the threshold for the hypernymy relations.

Computing Metrics based on a *term-term* matrix

Since other authors Caraballo (1999); Cederberg and Widdows (2003); Wandmacher et al. (2007)[6] propose computing LSA based on a *term-term* matrix $M(n, n)$, where n[7] is the total number of terms. And each entry, M_{ij} is the number of times, that

[6]The two last authors, use LSA for performing tasks very close to ours.
[7]$n \in \mathbb{N}$.

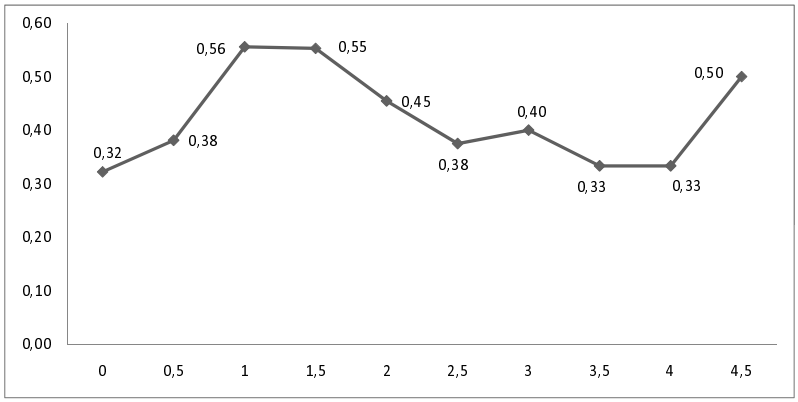
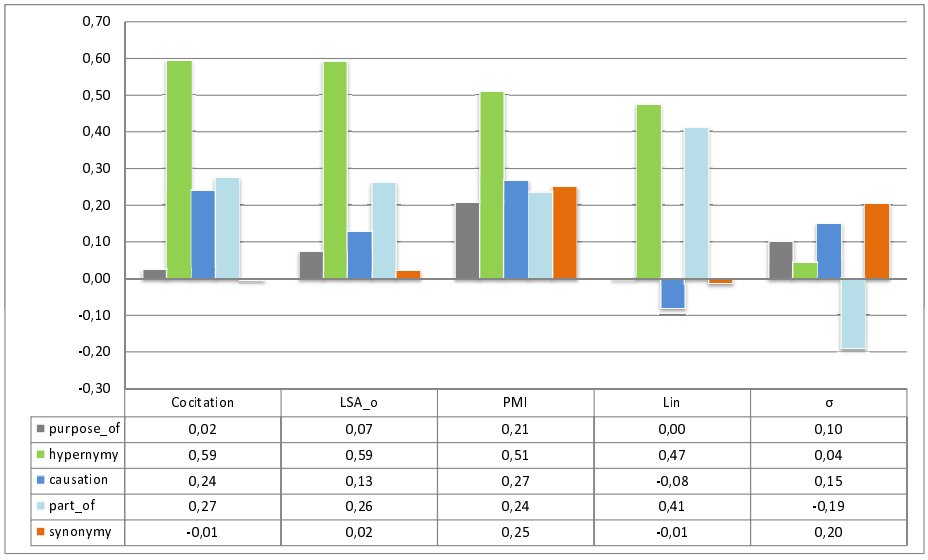


Figure 4.7: Evolution of the precision when increasing the threshold for the part_of relation.



	Cocitation	LSA_o	PMI	Lin	σ
purpose_of	0,02	0,07	0,21	0,00	0,10
hypernymy	0,59	0,59	0,51	0,47	0,04
causation	0,24	0,13	0,27	-0,08	0,15
part_of	0,27	0,26	0,24	0,41	-0,19
synonymy	-0,01	0,02	0,25	-0,01	0,20

Figure 4.8: Correlation coefficients between manual evaluation and the distributional metrics (*term-term* matrix).

terms i and j co-occur in a word context window. In this experiment, the word context window will be the bounded by the beginning and the end of a sentence.

To perform this experiment, the first 28,000 documents of CETEMPúblico, containing 30,100 unique content words (considering only nouns, verbs and adjectives), are analysed. Being the same documents used in the *term-document* matrix, result, in this case, in approximately 2.1 million *term-term* relations.

In figure 4.8, the correlation between the metrics and manual evaluation, is presented. The triples used in this experiment are the same used in the last experiments. The correlation values between the metrics and the manual validation in this experiment, (figure 4.8) are very similar to the correlation values obtained with the *term-document* matrix experiment (figure 4.3). So, in the next topic we will present a method to identify the correlation coefficient relevance between *term-document* and *term-term* values.

	Cocitation	**LSA_o**	**PMI**	**Lin**	σ
R_{1_h}	0.59	0.58	0.52	***0.42***	***0.04***
R_{1_p}	0.37	0.29	0.25	0.38	-0.19
$\varepsilon = 0.89$					
A_{2_h}	0.52	0.52	0.45	***0.41***	***0.03***
A_{2_p}	0.24	0.23	0.21	0.36	-0.16

Table 4.5: Statistical Dominance E_1 in E_2.

	Cocitation	**LSA_o**	**PMI**	**Lin**	σ
R_{2_h}	0.59	0.59	0.51	0.47	0.04
R_{2_p}	***0.27***	0.26	0.24	0.41	-0.19
$\varepsilon = 0.72$					
A_{1_h}	0.42	0.41	0.37	0.30	0.02
A_{1_p}	***0.26***	0.21	0.18	0.27	-0.13

Table 4.6: Statistical Dominance E_2 in E_1.

Measuring the correlation coefficient relevance between *term-document* matrix and *term-term* matrix

Currently, the comparison between algorithms for multi-purpose problems is an area that still has a lot to be done, however different indicators and methodologies have been proposed recently. A review of such methods is presented by Knowles et al. (2006), where the authors also suggest a methodology to test algorithms created to solve problems with multiple purposes.

In this work, a unitary additive epsilon (Zitzler et al. (2003)) is used in order to measure the statistical relevance of the correlation coefficient between the two last experiments (*term-document* matrix vs. *term-term* matrix). The unitary additive epsilon indicator is calculated using a reference set R.

Let E_1 be the set of hypernymy (R_{1_h}) and part_of (R_{1_p}) correlation values, ($R_{1_h}, R_{1_p} \in E_1$) which resulted from the experiment *term-document* (see section 4.1.2), and E_2 is the set of hypernymy (R_{2_h}) and part_of (R_{2_p}) correlation values, ($R_{2_h}, R_{2_p} \in E_2$) which resulted from the experiment *term-term* matrix (described in the last topic), where $R_{i_k} = \{Cocitation_{i_h}, LSA\_o_{i_h}, PMI_{i_h}, Lin_{i_h}, \sigma_{i_h}\}$, were $R_{i_k} \in \mathbb{R}$: $0 \leq R_{i_k} \leq 100$, i is triple number in the set T and $_k$ is the initial letter of the semantic relation name, i.e. $_k \in [_h, _p]$[8]. In this experiment, we only used hypernymy and part_of relations, because they have the greater correlation coefficient values, being that way the most relevant in ours experiment.

In order to calculate the unitary epsilon indicator for a set of approximation A_{j_k} from a reference set R_{i_k}, ε (epsilon) will be the value that must be multiplied to each reference R_{j_k}, resulting in a set weakly dominated A_{j_k} by the set R_{i_k}, i.e. $A_{j_k} = \varepsilon * R_{j_k}$, were $A_{j_k} \in \mathbb{R}$.

The results from E_1 experiment dominated the E_2 on 89% (see table 4.5), on the other hand, E_2 experiment dominated the experiment E_1 in 72% (see table 4.6). Based on this results the E_1 (*term-document*) experience has a greater statistical dominance comparing with E_2 (*term-term*).

[8] $_h, _p$ is the initial letter for hypernymy and part_of relations, respectively.

Manual Evaluation vs. Web Distributional Metrics

The Web distributional metrics presented in section 2.5.2, more precisely WebJaccard, WebOverlap, WebDice, WebPMI and WebNWD, equations 2.11, 2.12, 2.13, 2.14 and 2.15, respectively, assert the semantic similarity between any two entities, by giving a confidence value to the similarity between them. Based on this, we will use this metrics to study if they can be applied in knowledge validation, more specifically, in the triple validation task.

Based in Oliveira (2009) idea, we will use the aforementioned metrics with a set of indicative patterns I_x, proposed by us, to the five semantic relations, where $_x \in [_h, _p, _s, _c, _f]$[9]. Table 4.7 shows some of those patterns.

Semantic Relation	Indicative pattern (I_x)
Hypernymy	é|são um|uma
Synonymy	também conhecido|conhecida|chamado|chamada|designado|designada de|por|pela
Part_of	tem|possui|engloba|abrange|inclui|têm um|uma|vários|alguns
Causation	devido|derivado|derivada|causado|causada|resultado|efeito|consequência a|ao|á|por|pelo|pela| de|do|da
Purpose	usado|usada|utilizado|utilizada|através|objectivo|finalidade|intuito|serve no|na|para|de o|a|um|uma

Table 4.7: Semantic relations and their indicative patterns.

For this purpose, the previews aforementioned metrics need to be specified, namely in the search engine format, in our case, we have used the *Google Search API*[10] engine.

For instance, if the system extracts the triple $t_i = (e_1, r, e_2)$, we define three different queries, $P(e_1), P(e_2)$ and $P(e_1 \cap e_2)$, where:

- $P(e_1)$ is the number of search engine results for the query: {"e_1" + "I_x" + *}.
- $P(e_2)$ is the number of search engine results for the query: {* + "I_x" + "e_2"}.
- $P(e_1 \cap e_2)$ is the number of search engine results for the query: {"e_1" + "I_x" + "e_2"}.

Here, * represents a Web search engine wildcard that matches any potential word. By putting double quotes ("") around a set of words, the Web search engine consider the exact words in that exact order without any change. Attaching a + between a set of words, the Web search engine search for all the arguments precisely as we typed it.

The set of triples used in this experiment is the same used in the *Manual Validation* topic.

Each distributional metric was applied to the triple set T, in the following manner: for each triple $t_i = (e_1, r, e_2)$, the distributional similarity between $P(e_1)$ and $P(e_2)$ was computed. For multi-word entities, the metrics were applied between each word of one entity and each word of the other, in order to calculate the average value.

For example if we have the triple *personal computer* HYPERNYM_OF *laptop*:

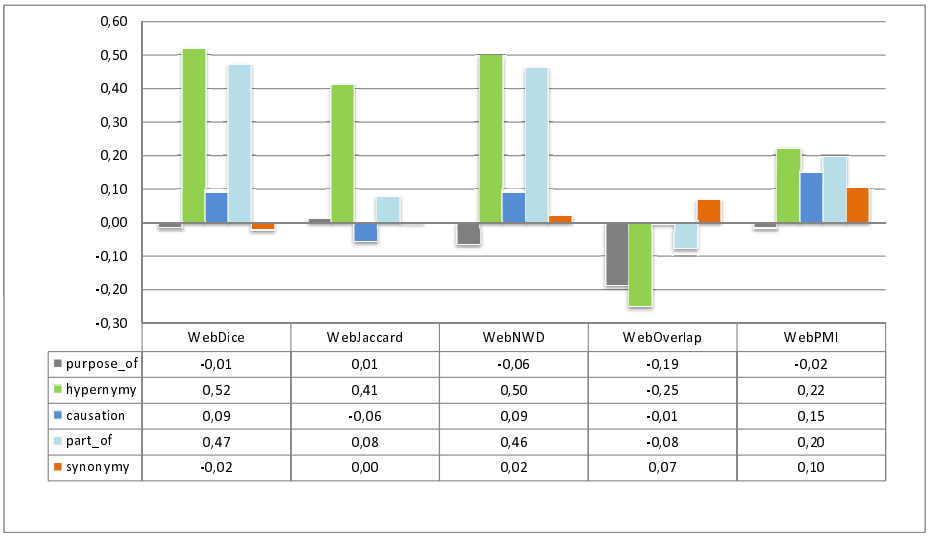
[9] $_h, _p, _s, _c, _f$ is the initial letter for hypernymy, part_of, causation, synonymy and purpose relation, respectively.

[10] http://code.google.com/intl/pt-PT/apis/ajaxsearch/web.html

- $P(e_1) = \frac{P(e_1)_1 + P(e_1)_2}{2}$, where $P(e_1)_1$ = "personal" + "I_h" and $P(e_1)_2$ = "computer" + "I_h";
- $P(e_2)$ = "I_h" + "laptop";
- $P(e_1 \cap e_2) = \frac{P(e_1 \cap e_2)_1 + P(e_1 \cap e_2)_2}{2}$, where $P(e_1 \cap e_2)_1$ = "personal" + "I_h" + "laptop" and $P(e_1 \cap e_2)_2$ = "computer" + "I_h" + "laptop".

For this purpose, the correlation coefficient presented in equation 2.16, section 2.5.3, will be used like in *Manual Evaluation vs. Distributional Metrics* topic of this actual section.

It is possible to observe in figure 4.9, that WebNWD, WebJaccard and WebDice metrics are strongly correlated with the hypernymy triples, and WebNWD and WebDice metrics for part_of triples, which suggests that hyponyms and their hypernyms, such as holonyms and their meronyms, tend to co-occur more frequently than causes and purposes.



	WebDice	WebJaccard	WebNWD	WebOverlap	WebPMI
purpose_of	-0,01	0,01	-0,06	-0,19	-0,02
hypernymy	0,52	0,41	0,50	-0,25	0,22
causation	0,09	-0,06	0,09	-0,01	0,15
part_of	0,47	0,08	0,46	-0,08	0,20
synonymy	-0,02	0,00	0,02	0,07	0,10

Figure 4.9: Correlation coefficients between manual evaluation and the Web distributional metrics.

To synonymy entities, despite sharing very similar neighbourhoods, may not co-occur frequently in the same sentence (Dorow (2006)) or even in the same document because they are alternative names for the same thing. For causation and purpose entities, we think that share some similar neighbourhoods, however may not co-occur frequently in the same sentence. Also, there are more ways to denote the purpose relations in corpora text which led to less extracted and more incorrect triples.

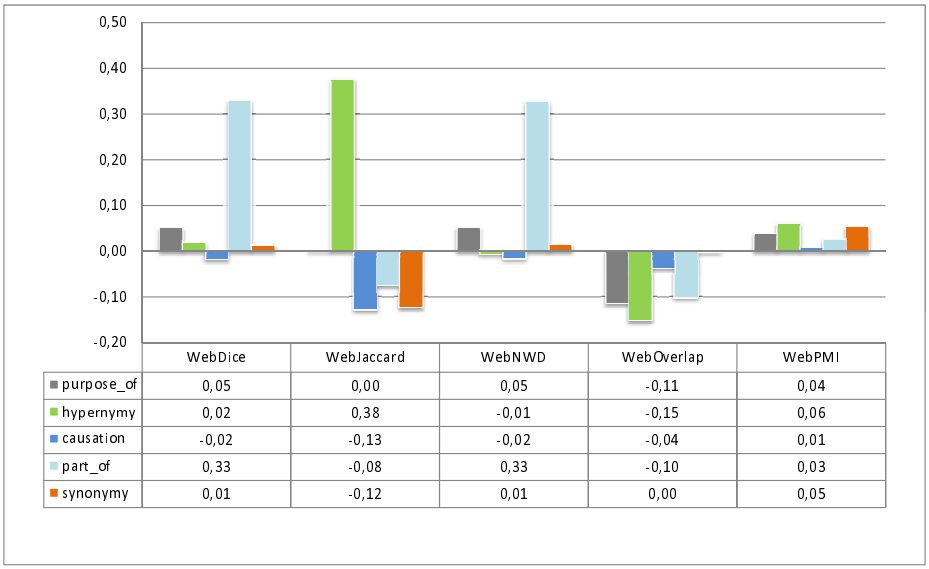
Although, another problem are the indicative patterns I_x used. In many queries the search engine return zero results, this might also be the reason for the low correlation coefficients for the last aforementioned semantic relations. We think this is understandable, because semantic terms can co-occur in many ways, or, in other words, each semantic relation can be translated in an enormous quantity of textual patterns, and we used just one per semantic relation, see table 4.7.

More limitations are related with the *Google Search API*, which is not sufficiently versatile to support a large number of expressions. Besides that, by searching a flexed term, Google is not capable of search terms that have the same lemma, which limits this kind of search.

An additional experiment, similar to the last one, but performed with complete entities, i.e. for each triple $t_i = (e_1, r, e_2)$, for example *personal computer* HYPERNYMY_OF *laptop*, the distributional similarity between $P(e_1)$ and $P(e_2)$ was computed in the follow manner:

- $P(e_1)$ = "personal computer" + "I_h" + *.
- $P(e_2)$ = * + "I_h" + "laptop".
- $P(e_1 \cap e_2)$ = "personal computer" + "I_h" + "laptop".

Figure 4.10, presents the results obtained.



	WebDice	WebJaccard	WebNWD	WebOverlap	WebPMI
purpose_of	0,05	0,00	0,05	-0,11	0,04
hypernymy	0,02	0,38	-0,01	-0,15	0,06
causation	-0,02	-0,13	-0,02	-0,04	0,01
part_of	0,33	-0,08	0,33	-0,10	0,03
synonymy	0,01	-0,12	0,01	0,00	0,05

Figure 4.10: Correlation coefficients between manual evaluation and the Web distributional metrics (with complete entities).

Despite 0.074% less in average for part_of, this experiment reinforce the idea that part_of entities tend to co-occur frequently. For hypernymy relation the WebNWD and WebDice correlation results are close to 0, however the correlation value for the WebJaccard decreased only 3%. That can indicate, that WebJaccard is the better option to validate triples with simple or multi words in their entities, achieving good results in both cases.

4.1.3 Conclusions

With this experiment, we have shown that the precision of systems capable of acquiring semantic knowledge from text, may benefit from applying distributional metrics to their output. Although this work is made for Portuguese, we believe that it can be adapted to other languages with similar distributional behaviour.

If, on the one hand, it is possible to combine several metrics in a linear expression or in a decision tree, on the other hand, the best results were obtained using an isotonic regression that selected the metrics which minimised the squared error.

Most of the works similar to ours, but for English, propose using LSA-based filters. However, despite very close correlation results, for hypernymy and part_of relations, our adaptation of the Cocitation metric, which is basically the Jaccard coefficient, seems to be the most adequate for such a task. Inspired by Cederberg and Widdows (2003) and Wandmacher et al. (2007) work, we have computed LSA based on a *term-term* matrix, nevertheless, in our experiment, *term-document* approach dominate the *term-term* matrix in 89%. However in a different scenario that can not be truth. We can not conclude that one approach is better than the other.

Furthermore, we have presented one way to validate triples based on search engine hits. Besides, high correlation values for part_of and hypernymy, the results for the others semantic relations are not relevant.

We think this is understandable, because semantic terms can co-occur in many ways, and each semantic relation can be translated in an enormous quantity of textual patterns.

4.2 Experiment 2: knowledge extraction from Wikipedia

In this section, we present our improved system. More specifically, we have improved the **Knowledge Extraction** module, presented in section 3.2, to perform the automatic extraction of relational triples extracted from the Wikipedia abstracts. There have been improvements at three levels: (1) more grammar patterns; (2) POS tagger integration to identify adjectives and to enable the next point; (3) extraction of hypernymy relations from multi-word terms, described in section 3.2.3.

We start by presenting the experimentation goals, and all the modules used to perform it, and then, the results obtained using our methodology on the Wikipedia abstracts. In the end, some limitations of this work are discussed.

4.2.1 Experiment Goals

Encyclopedias are classified as semi-structured resources, because they have entries for different entities, and besides that, the encyclopedia's content is not only about words; includes knowledge about the world and human knowledge. Caused as well by its availability on the Web, in the last few years the use of encyclopedias, such as Wikipedia, became more and more used by the scientific community in different tasks, such as information extraction and information retrieval.

Having in mind its collaborative construction, this resource is an endless information source in constant evolution. So, we have decide to use the Portuguese version of the Wikipedia abstracts[11] in our system, in order to automatically extract semantic knowledge for it.

Comparatively to the system used in the CETEMPúblico experiment, see section 4.1, in this experiment we will use an improved version of our grammars, and

[11]The Wikipedia abstracts describe in few words the article's content, and so, they have the most important information and less variations in its construction.

consequently it will improve our system. How the used textual patterns performs in the extraction process, will be study too.

Additionally, we will study one alternative to automatically evaluate semantic knowledge, extracted automatically from the abstracts.

4.2.2 Experiment

Experiment Set-up

Figure 4.11, describes all the modules used over the abstracts of the Portuguese version of Wikipedia.

This approach is based in Portuguese patterns, that are indicators of semantic relations in text, like is referred in the chapter 3. Some of them can be found in table 4.11.

System set-up: The **Knowledge extraction** topic, presented in figure 4.11, were prepared to analyse text, one sentence at a time. So, the first phase prepares the abstracts, separating into sentences, **Text Splitter** module, in order to prepare it for the extraction module, where it is processed according to the grammars, **Automatic extraction of relational triples** module.

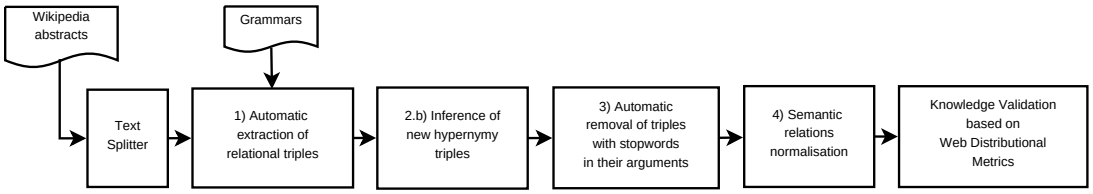


Figure 4.11: Modules used in *Experiment 2*.

In this system version, we also chose to extract semantic relations between multi-word terms in text, and complex entities, taking advantage of two lexical-syntactic patterns: [**N ADJ|ADV**] and [**N de|do|da|com|para N**], **2(b) Inference of new triples** module - the 2(b) method is described in section 3.2. For example, if a term occur modified by an adjective (e.g. *Computador pessoal*; in English *personal computer*), or by a preposition after and before a noun (e.g. *Garrafa de água*; in English *Bottle of water*), a new triple will be inferred *Garrafa* HYPERNYM_OF *Garrafa de água*[12], see section 3.2.3 for more details.

In order to identify the grammatical categories of the words, it is previously realized a syntactic analysis of each sentence, using a module for the POS tagger included in the OpenNLP2 package, section 2.6. This module was trained with Bosque, section 2.4.1.

The grammars contain essentially lexical patterns, and rely on syntactic analysis just only to identify adjectives. Besides the morfo-syntactic analysis, every word in the sentence is lemmatised, using other OpenNLP module, to which was added a set of regular rules to transform plural into singular.

After, triples whose arguments are in a list of not desired words, (essentially stopwords) are removed, **Automatic removal of triples with stopwords in their arguments** module in the figure 4.11.

[12] *Bottle* HYPERNYM_OF *Bottle of water*

At last, the name of the semantic relation is changed based on the grammatical category of its arguments, **Semantic relations normalisation** module, as table 3.2 in section 3.2 describes.

Wikipedia set-up: We soon verified that almost all of Wikipedia content were very specific, and useless in the construction of a lexical ontology, containing articles about personalities, organisations or historical eras. Due to this problem, we searched a solution to filter abstracts associated with Named Entity Recognition (NER), which could lead also to a decrease of the amount of text that needs to be processed.

To do that, we use the abstracts made available by DBPedia (Bizer et al. (2009)) and their taxonomy. Besides that, DBPedia maps the Wikipedia in a taxonomy where each article receives one or more high level types, like *Person*, *Place*, *Organisation*, *Mean* of *Transportation*, *Device* or *Species*, and some more specific types like *Writer*, *Airport*, *Soccer Club*, *Bird*, *Automobile* or *Weapon*.

Although the attribution of types is not available for the Portuguese version of Wikipedia, there is a correspondence between the identifiers of the entries from the various Wikipedia's languages, that have inserted the same subject. Taking that in consideration, we use the types assigned to the entries of the English version to filter entries from the Portuguese version with the types *Person*, *Place*, *Organisation*, *Event*, and others that are associated to NER category.

Nevertheless, there are many entries from the Portuguese Wikipedia that do not have correspondence, however there was possible to eliminate a large quantity of entries related to NER. More precisely, we removed 30% from a total (368,521 abstracts), resulting in 494,187 abstracts[13], that we will call as set **A1** .

Even so, we still had a large quantity of text that was useless, such as entries about Portuguese and Brazilian geography.

Because of that, although the fact that we could lose interesting entries that only exist in the Portuguese version of Wikipedia, we chose to decrease the amount of abstracts, and kept only entries that, through the taxonomy, we could confirm that belong the following types: *Species*, *Anatomical Structure*, *Chemical Compound*, *Disease*, *Currency*, *Drug*, *Activity*, *Language*, *Music Genre*, *Colour*, *Ethnic Group* and *Protein*.

Besides that, although some abstracts consist of two or three sentences, we chose to use just the first sentence of each abstract, because its contains the most relevant information of the abstract. So, we got a total of 37,898 sentences to process that we will call as set **A2**, which we consider the most important on our experiment.

Experiment Results

While the extraction from the set **A1** set took about 12 days to run in a machine with a 1.6GHz dual core processor and a virtual machine with a 256MB heap, the processing of the set **A2** set only took about 1 day.

The quantity of triples extracted from both sets (**A1** and **A2**), before (Total) and after removing (A/rem) repeated triples, is presented in table 4.8, where the number of triples of which set are put side-by-side, with some examples.

[13]Some abstracts consist of two or three sentences, however we chose to use just the first sentence of each abstract. Usually, it contains the most relevant information of the abstract.

For hypernymy, we separate the triples extracted based on the **2(b) Inference of new hypernymy triples** module (Hypernymy *Inf*[14]) analysis, from the triples extracted trough the grammars.

Relation	**Extracted-A1**		**Extracted-A2**		**Examples**
	Total	A/rem	Total	A/rem	
Hypernymy *Inf*	711,954	390,492	24,367	16,228	(*desordem*,*desordem_cerebral*) (*átomo*,*átomo_de_carbono*)
Hypernymy	149,845	144,839	31,254	29,563	(*desporto*,*automobilismo*) (*estilo_de_música*,*folk*)
Synonymy	25,816	25,518	11,872	11,862	(*inglês_antigo*,*anglo-saxão*) (*estupro*,*violação*)
Part_of	12,093	11,485	1,321	1,287	(*jejuno*,*instestino*) (*rolas*,*columbidae*)
Purpose	13,277	12,992	777	743	(*amoxicilina*,*tratamento_de_infecções*) (*construção*, *terracota*)
Causation	5,854	5,740	559	520	(*parasita*, *doença*) (*doença_neuromuscular*,*fadiga*)

Table 4.8: Number of triples extracted from the Wikipedia abstracts.

After analysing the results of the set **A2**, it is possible to observe that is extracted almost one hypernymy relation per sentence, through the analysis of textual patterns.

This occurs because many abstracts begin with the construction [X *é um* Y], resulting in X HIPERNYM_OF Y.

Besides that, there are sentences that have an enumeration in the place of the hypernym X, which originates one hypernymy relation for each enumerated term.

For example, the sentence S(7), originates four triples: t_1, t_2, t_3 and t_4.

S(7): *A heroína ou diacetilmorfina é uma droga.*[15]

t_1: *droga* HYPERNYM_OF *heroína*;

t_2: *droga* HYPERNYM_OF *diacetilmorfina*;

t_3: *heroína* SYNONYM_OF *diacetilmorfina*;

t_4: *diacetilmorfina* SYNONYM_OF *heroína*.

Another curiosity, in both sets, is related to the hypernymy relations, that relate essentially plants, animals, or others living creatures to a specie, order or class entities. The purpose relations normally associate health problems to their therapeutics, and causation relations are as well established between health problems, associating their causes and effects. The synonymy relations are normally established between terms in their European variant of Portuguese and the Brazilian one, for example, *marrom* SYNONYM_OF *castanho*[16] or *esófago* SYNONYM_OF *esôfago*[17].

[14]The Hypernymy *Inf* triples were extracted from multi-words, see *Extracting Triples based on Multi-Word Terms* topic in section 3.2.3 for more details.

[15]*Heroin or diacetylmorphine is a drug.*

[16]Similar to *brown* SYNONYM_OF *brunet*.

[17]The terms are the same, but one of them are written in Brazilian Portuguese, and the other in European Portuguese. Similar to the triple: *lemmatisation* SYNONYM_OF *lemmatisation*, where one of the terms is written in European English and the other in American English.

Besides that, the sentences from which are extracted synonymy relations, most part of them begin with a large quantity of synonym enumerations. The extreme example is the sentence that starts by: "*Bagre-bandeira, bagre-cacumo, bagre-de-penacho, bagre-do-mar, bagre-fita, bagre-mandim, bagre-sari, bandeira, bandeirado, bandim, pirá-bandeira, sarassará, sargento or bagre-bandeirado... é um peixe da família dos ariídeos...*"[18].

Manual Evaluation

In order to evaluate the knowledge extracted by our system, we have used the scale proposed by Freitas (2007), which suggests four groups to classify triples:

- **0**, if the triple is completely incorrect;
- **1**, if the triple is correct, but too general or specific to be useful;
- **2**, if the triple have a preposition or an adjective that makes one of the arguments strange and prevents the triple from being correct;
- **3**, if the triple is correct.

Initially, there were generated 12 random samples containing ≈80 triples each, taken from the set **A1**, and each of them was manually evaluated by two reviewers. Table 4.9 presents the results obtained, and also the exact agreement (EA) and the relaxed agreement (RA) between them. In the relaxed agreement, we have considered the values 1 and 3 as correct, and 0 and 2 as incorrect.

Relation	**Evaluated**	**3(%)**	**2(%)**	**1(%)**	**0(%)**	**EA(%)**	**RA(%)**
Hypernymy	240	34	6	48	12	53	85
Synonymy	199	70	9	6	15	76	88
Part_of	182	23	23	16	38	54	74
Purpose	158	26	14	9	51	61	78
Causation	183	22	21	48	9	52	73

Table 4.9: Manual evaluation results of the set **A1**.

However, in this evaluation we verified that many triples were useless, see *Wikipedia set-up* topic, to their use in a lexical ontology, like those that indicate geographical sub-divisions (e.g. *sub-região estatística portuguesa* PART_OF *região do alentejo*), those that are related with historical eras (e.g. tragédia de 1892 CAUSATION_OF crise política), among others too specific (e.g. *romancista brasileiro* PART_OF *academia brasileira de letras*, *escola* HYPERNYM_OF *escola de música Juiliard*).

Besides that, it was verified that, essentially due to the POS tagger limitations, and also due to the type of the text processed, most part of the triples which were changed in the final phase, **Semantic relations normalisation** module, associated to the fact that one or both of its arguments were not nouns, the triple normalisation was incorrect performed. So, we chose to continue the evaluation using only relations between nouns.

[18] http://pt.wikipedia.org/wiki/Bagre-cacumo

Taking all of that in account, new triples were selected for manual evaluation, more specifically 12 random samples were taken from the set **A2**, with 87 triples each. Once again, each sample was reviewed by two reviewers. We also have used triples classified in the last evaluation, set **A1**, where 663 are presented in the set **A2**, making a total of 1707 triples. In order to analyse the quality of the triples, the reviewers were advised to search the Web, including Wikipedia itself, for information about the entities that were involved.

The results of this second evaluation can be found on table 4.10, where the values presented are the add of both reviewers classifications. The table, also contains the exact agreement (EA) and the relaxed agreement (RA) between them, and where we have considered, the values 1 and 3 as correct, and 0 and 2 as incorrect.

The relaxed agreement (RA) was calculated because is not always easy to identify if a triple is too general or too specific to be useful to a lexical ontology. Besides that, this Knowledge could be used in the future in other purpose and, apart from that, the triples classified with 1 are also 'correct'.

Relation	**Evaluated**	**3(%)**	**2(%)**	**1(%)**	**0(%)**	**EA(%)**	**RA(%)**
Hypernymy *Inf*	323	35.0	4.2	42.1	18.7	57.3	82.7
Hypernymy	322	57.5	33.8	1.6	7.1	89.8	93.1
Synonymy	286	85.7	7.3	0.4	6.6	90.0	91.6
Part_of	268	44.2	26.7	8.4	20.7	63.1	78.4
Purpose	264	53.0	16.5	4.0	26.5	71.2	82.2
Causation	244	41.8	24.6	7.8	25.8	61.5	79.5

Table 4.10: Manual evaluation results of the set **A2**.

An interesting point, when we compare the results of both sets, is the difference in the number of triples classified with 1, which is a lot bigger in the set **A1**. In proportion, this number decreased 39% (set **A1**) to 22% (set **A2**) from the total. Also in proportion, we report an increase of correct triples, mainly in causation and purpose relations, where the number increased 1,5 and 2 times more, respectively. Concordances have also increased slightly. These improvements in the set **A2** are consequence of a more limited group of entries, and where exists less ambiguity. Still, about one quarter of the triple of causation and purpose and a fifth part of the triple still completely wrong, which is essentially related to the ambiguity of some patterns used.

In table 4.10 we can notice a greater agreement on the division between correct and incorrect triple, EA and RA respectively, mainly because it embraces a more objective division, where do not enters the subjectivity of evaluating the actual usefulness of a triple in a lexical ontology. For example, several hypernymy triple, extracted using the **Inference of new hypernymy triples** module, do not add much significant information to a lexical ontology (e.g. *equipa* HYPERNYM_OF *equipa de seis jogadores*[19]), however this classification is very sensitive to the discretion of the reviewers.

This happens because the proportion of sentences of species increased in the set **A2**, and many of these species are identified by two words. For example, in the sentence "*O Iriatherina werneri é uma espécie de peixe de aquário*"[20], the POS tagger does not identify the two words as one entity *Iriatherina werneri*, which leads

[19] *team* HYPERNYM_OF *team of six players*

[20] *The Iriatherina werneri is a species of aquarium fish.*

the system to not interpret the entity as a modified noun, and therefore, extracts a triple with an incomplete argument, *peixe de aquário* HYPERNYM_OF *werneri*, instead of *peixe de aquário* HYPERNYM_OF *Iriatherina werneri*[21].

Studying Patterns Efficiency

Besides the evaluation of the quality of the extracted triples, was performed a study about the patterns or key-words that originated more triples. For the set **A2**, that information can be found on table 4.11. To that we additionally added information about the classification obtained in the manual evaluation by the triples extracted through these patterns. In this case, we only consider the triples that had the same evaluation value by both reviewers.

Relation	Pattern	Extracted	Evaluated			
			3	2	1	0
Hypernymy	*multi-word term*	24.367	72	7	75	32
	é uma espécie de	15.824	54	96	0	0
	é um|uma	10.960	87	11	0	15
	é um género de	2.402	24	0	0	0
Synonymy	ou	4.886	154	2	0	2
	também conhecido|a|os|as por|como	3.016	60	4	0	4
Part_of	inclui|incluem	471	34	0	2	15
	grupo de	158	17	3	1	0
Purpose	utilizado|a|os|as para|como|em|no|na	376	71	16	1	20
	usado|a|os|as para|como|em|no|na	237	41	3	1	4
Causation	causado|a|os|as	165	27	11	1	10

Table 4.11: Quantity of triples extracted based on their indicative patterns.

From the patterns that lead to a more incorrect triples in the extraction, we highlight [**usado**|**utilizado**] that, when followed by [**em**|**no**|**na**] can indicate not the purpose relation, but the spot where an object is used, like in "*O Ariary malgaxe é a moeda usada em Madagáscar*"[22]. Another pattern that appears to be quite ambiguous is: [**inclui**|**incluem**]. On the other hand, the pattern [**é um género de**], only resulted in correct hypernymy triple.

Validating triples using the CETEMPúblico *term-document* matrix

In this experiment, we will use the *term-document* matrix, to validate triples extract from the Wikipedia abstracts. It was created with the first 28,000 documents of CETEMPúblico corpus, the same matrix[23] used in the section 4.1.

Since the triples were extracted from the Wikipedia abstracts and the corpus used to weighting them is different, we had to, at first, select triples that have there words in the *term-document* matrix. These numbers are presented in table 4.12 (column Total).

To evaluate the precision of the results, we selected random samples for each type of relation. The samples' sizes took the type of relation into consideration and they were divided into two random samples, each one evaluated by one human

[21] *aquarium fish* HYPERNYM_OF *Iriatherina werneri*

[22] *The malagasy ariary is the currency of Madagascar.*

[23] See section 4.1 to understand this *term-document* matrix construction.

judges. Each human judge was asked to assign one of the following values[24] to each triple, according to its quality:

- **0**, if the triple is completely incorrect.
- **1**, if the triple is not incorrect, but something is missing. Like a preposition or an adjective that makes one of the arguments strange and prevents the triple from being correct in one or both of its arguments, or even the relation is very generic.
- **2**, if the triple is correct.

The results obtained for manual evaluation are described in the table 4.12. Column Evaluated presents the number of triple manual evaluated, and the others columns present the percentage of triples manual evaluated with 0, 1 and 2.

Relation	**Total**	**Evaluated**	**2(%)**	**1(%)**	**0(%)**
Hypernymy *Inf*	2,211	346 (15.6%)	17.7	64.7	17.6
Hypernymy	702	437 (62.2%)	55.5	14.6	29.9
Synonymy	592	391 (66%)	54.8	5.6	39.6
Part_of	55	268 (100%)	21.8	20	58.2
Purpose	57	264 (94.7%)	31.5	20.4	48.1
Causation	98	24 (84.7%)	33.8	21.7	44.5
TOTAL	***3,715***	***1,760 (47.4%)***			

Table 4.12: Manual evaluation results of triples with their entities in the CETEMPúblico *term-document* matrix.

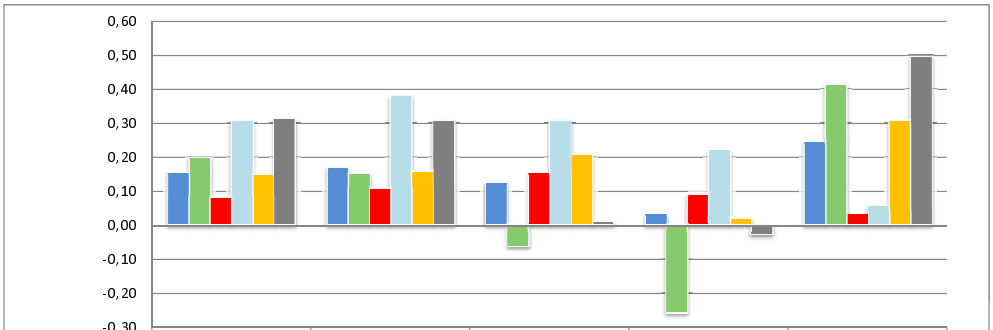
As we can see, the manual evaluation percentage given to the triples, table 4.12, is little different to the results presented in table 4.10. However, this can be explained by the scale used (0, 1 and 2), raising ambiguity in some cases, an leading to more triples evaluated with 0, that with the scale used in the *Manual Evaluation* topic in this section (0, 1, 2 and 3), some of them would be classified with 1. Besides that, the number of triples that have all of their words in the *term-document* matrix is scarce, concretely for part_of, purpose and causation, 55, 57 and 98 respectively.

Another curiosity that have leaded to a less triples classified with 0, is related to the fact that in this set of triples, the entity *ser* appears 118 times. Ninety-nine of them were classified with 0 (e.g. *cloreto* SYNONYM_OF *ser*, *germânico comum* HYPERNYM_OF *ser*, *língua* SYNONYM_OF *ser*, *ser* PART_OF família; in English *chloride* SYNONYM_OF *be*, *common germanic* HYPERNYM_OF *be*, *language* SYNONYM_OF *be*, *be* PART_OF family). Once again due to the POS tagger limitations, it classifies the verb 'to be' wrongly, even so, the text processed rises some difficulties too.

Nevertheless, in this experiment we are not so interested in the manual evaluation values, instead we are more interested in the possibility of automatically validation these knowledge with the metrics presented in section 2.5.1, and prove their application in the triples validation task - like we have done, with a different text in *Experiment 1*, section 4.1.

So, in order to study the relationships between the manual evaluation and the values given by the metrics, the correlation coefficients between them were computed. Figure 4.12 shows these correlation coefficients values.

[24]The values used in this experiment were the same used in the *Experiment 1*, section 4.1.



	Cocitation	LSA_o	PMI	Lin	σ
hypernymy	0,16	0,17	0,13	0,03	0,25
purpose_of	0,20	0,15	-0,06	-0,26	0,42
hypernymy_inf	0,08	0,11	0,15	0,09	0,03
causation	0,31	0,38	0,31	0,23	0,06
part_of	0,15	0,16	0,21	0,02	0,31
synonymy	0,31	0,31	0,01	-0,03	0,50

Figure 4.12: Correlation coefficients between manual evaluation and the distributional metrics.

It is possible to observe that most of the metrics are mightily correlated with the quality of the triples, except for hypernymy. This may happen because the set of triples used, in this experiment, are associated essentially with plants, animals, or others living creatures to a specie, order or class entities, and in the *term-document* matrix, created from a journalistic corpus, these entities do not appear frequently, leading to a low correlation values. In particular to the inferred triples (Hypernymy *Inf*) the same explanation can be taken, because their entities represent general or specific knowledge, appearing scarcely in the journalistic corpus.

Another curiosity, is related to the purpose relations, that normally associates health problems to their therapeutics, and causation relations are as well established between health problems, associating their causes and effects. Despite that, their entities seems to occur frequently in the matrix, raising to 38% with LSA and 42% with σ values for causation and purpose, respectively.

Surprisingly, the weight correlation value is related to the synonymy relation with the σ metric. We think this can happen because the synonymy entities evaluated occurs frequently in the *term-document* matrix, leading to a weightier values with Cocitation, LSA and particularly to σ. However, for the last mentioned we can not conclude that metric is or not the better metric to validate synonymy triples, because it is based on the relevance of the words in the entities, and do not have in consideration their co-occurrences.

Furthermore, for this particular experiment the LSA metric gives weightier correlation values than Cocitation, different from the *Experiment 1*, nevertheless we do not think that one metric is better than the other. More specifically, we think that each one is more appropriated to a specific semantic relation, and also their application depends on the context. For example, Turney (2001) concluded that PMI performed better than LSA, measuring the similarity of pairs of words in synonymy relation, on the Test of English as a Foreign Language (TOEFL[25]). However,

[25]http://www.ets.org

other authors, such as Budiu et al. (2007) that have compared LSA, PMI and GLSA (Matveeva et al. (2005)), have concluded that LSA and GLSA (with essentially combines PMI with SVD[26]) outperform PMI in a small domain-specific corpus, such as TASA (Zero et al. (1995)), still PMI performed better in a large web-corpus than LSA.

4.2.3 Automatic Evaluation Proposal

It is known that manual evaluation is probably the most reliable form to evaluate semantic relations, however it is a slow and tedious task, and can be subjective, independently from the quantity of criteria to guide the reviews. This can be verified by looking to our manual evaluation agreement in table 4.10. Although we have used two different ways to measure the agreement, it can be hard to distinguish from the various classifications of a scale. For example, besides the subjectivity existing when we decide the utility of a triple, the distinction between 1 and 2, classification values, may be not too clear. Also, a triple can be too general, too specific, or one modifier can be missing raising subjectivity in the evaluation. Besides that, this kind of evaluation can not be repeated with an automatic evaluation approach, this would no longer be a problem. That led us to propose an automatic evaluation approach.

Set-up

One way to automatically evaluate knowledge resulted from information extraction - is to use the information available on the Web. In the case of semantic triples validation, an alternative would be to search sentences whose argument's relationship is explicit through textual patterns. This is done for example in Costa et al. (2010), however over a journalistic corpus.

Following these ideas, the automatic evaluation of the triples extracted from the Wikipedia abstracts, will have as base the application of the five metrics, normally used to measure the similarity between two entities (Bollegala et al. (2007)), in the web, more precisely: WebJaccard, WebOverlap, WebDice, WebPMI, and WebNWD, equations 2.11, 2.12, 2.13,2.14 and 2.15 respectively. All presented in section 2.5.2.

As referred above, this metrics are normally used to calculate the distributional similarity between two terms, based on their occurrence and neighbourhoods, and, although relational terms usually have similar distributions, these metrics do not have into account their semantic relation.

So, inspired by Oliveira (2009) idea, in order to apply these metrics to semantic triples validation, it should also be included a textual pattern that identifies the relation. Similar what we have done in the section 4.1, we will use a set of semantic patterns S_x, proposed and improved by us, to the five semantic relations, where $_x \in [_h, _p, _s, _c, _f]$[27].

Nevertheless, in text only a subset of relations are in the type X RELATION_OF Y, where X is the entity that contains Y, and Y the entity contained by X. Sometimes in text, X appears firstly then Y, other times Y appears firstly than X. So, in order to embrace all the possibilities, we propose semantic patterns for both

[26]SVD: Singular Value Decomposition.

[27]$_h, _p, _s, _c, _f$ are the initial letter for **h**ypernymy, **p**art_of, **s**ynonymy, **c**ausation, and **p**urpose relation respectively.

combinations. Table 4.13 shows these semantic patterns.

Semantic Triple	Semantic Patterns (S_x)
X HYPERNYM_OF Y	como|tipo|incluindo o|a|os|as|um(a|as)|uns **tal|tais como o|a|os|as|um(a|as)|uns** como por exemplo o|a|os|as|um(a|as)|uns
Y HYPONYM_OF X	**é|são um(a)** é|são um(a|as)|uns tipo(s)|forma(s)|classe(s)|género(s)| espécie(s)|variedade(s)|raça(s) de **e|ou|entre outros(as)|mais** e|ou ainda|também outros(as)|mais
X SYNONYM_OF Y *or* Y SYNONYM_OF X	**também conhecido(a|as|os)|chamado(a|as|os)|designad(a|as|os)** **de|por|pela** ou
X PART_OF Y	**tem|possui(em)|engloba(m)|abrange(m|inclui(em)|têm** **um(a|as)|vários(as)|alguns|uns** constituído(a|das|dos)|formado(a|das|dos)|composto(a|das|dos) de|por|pelos(as) um(a|as)|uns|algum|alguns **inclui(em)|abrange(m)|engloba(m) o|a|os|as|um(a|as)|uns** pode(m)|deve(m)|costuma(m) ter|possuir
Y HAS_PART X	**é|são membro(s)|elemento(s)|espécie(s)|porção(ões) de(o|a|as|os)** **grupo|conjunto|família|classe|clã de(o|a|as|os)** do(a) grupo|conjunto|família|classe|núcleo|clã de(o|a|as|os) **faz parte de(o|a|as|os)** faz parte de um(a|as)|uns **pertence(m) ao(s)|à(s)** é|são um ramo de(o|a) **inserido(a|as|os)|incluído(a|as|os) em|no(a|as|os)|num|nuns|numa(s)** é|são|está(ão) inserido(a|as|os)|incluído(a|as|os) em|no(a|as|os)|num|nuns|numa(as) **constitui(em)|compõe(m)|forma(m)|estabelece(m)|constitui(em)|** **compõe(m) o(a|as|os)|um(a|as)|uns** faz|fazem parte de(o|a|as|os)
X CAUSATION_OF Y	**causa(m)|origina(m)|resulta(m) em|no(a|as|os)** pode(m) causar|resultar|originar em|no(a|as|os) **pode(m) causar|resultar|originar** tem|têm como resultado(s)|efeito(s)|causa(s) **é|são causador(es) de(o|a|as|os)** resulta(m) em|no(a|as|os)|num|nuns|numa(as)
Y EFFECT_OF X	**resulta(m)|resultado de(o|a|as|os)** causado(a|as|os)|provocado(a|as|os)|derivado(a|as|os)| originado(a|as|os) por|pelo(a|as|os) **devido|derivado(a|as|os) a|ao(s)|à(s)**
X PURPOSE_OF Y	é|são objectivo|objectivos|finalidade **tem|têm como objectivo(s)|finalidade(s)** com o(s) objectivo(s)|finalidade de **serve(m) para** frequentemente utilizado(a|as|os) como **utilizado(a|as|os) como|em|para|no(a|as|os)** usado(a|as|os)|atravs|objectivo|finalidade|intuito| serve(m)|destina-se|destinam-se para|no(a|as|os)
Y MEANS_FOR X	**através de(o|a|as|os)**

Table 4.13: Semantic relations and their indicative textual patterns.

In order to use this patterns with the previews aforementioned metrics, they need to be specified, namely in the search engine format. In our case, we have used the *Yahoo Search Web Service*[28] engine. It does not support the Portuguese region in the parameter requests, so we have used the Brazilian region. We think this can not raise any problem, because Brazilian language is very similar to the Portuguese.

For instance, if the system extracts the triple $t_i = (e_1, r, e_2)$, we define three different queries, $P(e_1), P(e_2)$ and $P(e_1 \cap e_2)$, where:

- $P(e_1)$ is the number of search engine results for the query: $\{e_1$+NEAR+$S_x\}$.

[28]http://developer.yahoo.com/search - this API is limited to 5,000 queries per IP address per day and non-commercial use.

- $P(e_2)$ is the number of search engine results for the query: $\{S_x$+NEAR+$e_2\}$.
- $P(e_1 \cap e_2)$ is the number of search engine results for the query: $\{e_1$+NEAR+S_x+NEAR+$e_2\}$.

Using the word 'NEAR' between two words or phrases finds pages where those words co-occur close to one another in any order. The symbol '+' between words is necessary in the Yahoo Web Service REST call. Besides that, if an entity contains multi-words, it is necessary to add the word 'AND' between them, e.g. if e_1= *Personal computer*, the query $P(e_1)$ will be {Personal+AND+computer+NEAR+S_x}. Moreover, if e_1 is the entity that contains e_2: $P(e_1) = \{S_x$+NEAR+$e_1\}$, $P(e_2) = \{e_2$+NEAR+$S_x\}$ and $P(e_1 \cap e_2) = \{e_2$+NEAR+S_x+NEAR+$e_1\}$.

With $P(e_1)$, $P(e_2)$ and $P(e_1 \cap e_2)$, will be possible calculate all the Web metrics to the triple set T.

Manual Evaluation vs. Web Distributional Metrics

The first step was to calculate these metrics for each manually evaluated triple, whose classification was concordant for both reviewers (the percentage of these triples is presented in table 4.10, column EA(%)).

We calculated these metrics with the *Yahoo* returned hits, in the evaluated triples and with the patterns presented in table 4.13. We experimented two approaches: (i) using all the semantic patterns an make a weighted average, (ii) using the weightier semantic patterns returned. For example if we have the triple *planeta* HIPERONIMO_DE *marte*:

i) making a weighted average:

- $P(e_1 \cap e_2) = \frac{P(e_1 \cap e_2)_1 + P(e_1 \cap e_2)_2 + ... + P(e_1 \cap e_2)_n}{n}$;
- $P(e_1) = \frac{P(e_1)_1 + P(e_1)_2 + ... + P(e_1)_n}{n}$;
- $P(e_2) = \frac{P(e_2)_1 + P(e_2)_2 + ... + P(e_2)_2}{n}$;
  where, $P(e_1 \cap e_2)_1 = \{$"planeta" + "S_{h1}" + "marte"$\}$, ..., $P(e_1 \cap e_2)_n = \{$"planeta" + "S_{hn}" + "marte"$\}$;
  $P(e_1)_1 = \{$"planeta" + "S_{h1}"$\}$, ..., $P(e_1)_n = \{$"planeta" + "S_{hn}"$\}$;
  $P(e_2)_1 = \{$"S_{h1}" + "marte"$\}$, ..., $P(e_2)_n = \{$"S_{hn}" + "marte"$\}$.

ii) choosing the weightier hit S_h:

- $P(e_1 \cap e_2) = \max(P(e_1 \cap e_2)_1, P(e_2 \cap e_1)_2), ..., P(e_2 \cap e_1)_n)$;
- if $P(e_1 \cap e_2)_1 = \{$"marte" + "`é|são um(a)`" + "planeta"$\}$ is the foremost weighted:
  - $P(e_1) = \{$"marte" +"`é|são um(a)`"$\}$;
  - $P(e_2) = \{$"`é|são um(a)`" + "planeta"$\}$.

n is the number of semantic patterns in the set S_h.

The choosing the weightier hits S_h approach returned the better results. However, we verified that we obtained values only for a small quantity of triples (20% of the concordant), because the rest never co-occurred with the chosen pattern. Also we have used only the triples evaluated with 3, table 4.14 present these triples and their Web distributional metrics values (WJ -WebJaccard, WO -WebOverlap, WD -WebDice, WP -WebPMI and WN -WebNWD), and also the number of hits (Hits) returned.

Relation	Metrics (%)						Examples
	Hits	WJ	WO	WD	WP	WN	
Hyper. *Inf*	3740	≈0	≈0	0.8	0.9	28	(*síndrome,síndrome de McCuneAlbright*) (*syndrome,cCuneAlbright syndrome*)
Hypernymy	5	≈0	≈0	2.9	≈0	22.3	(*ave,tegimae*) (*bird,tegimae*)
Synonymy	7	≈0	≈0	≈0	≈0	≈0	(*baleia-glacial,baleia-sardinheira*) (*baleen whale,sei whale*)
Part_of	3	0.1	≈0	0.2	0.3	34.9	(*damão-do-cabo,ordem hyracoidea*) (*cape hyrax,order hyracoidea*)
Purpose	3	≈0	≈0	0.3	0.1	34.4	(*tratamento do impetigo,mupirocina*) (*impetigo treatment, mupirocin*)

Table 4.14: Example of triples and their Web distributional metrics values.

Still, the returned values of the metrics are far from the expected. This is understandable, because the semantic terms can co-occur in many ways, or, in other words, each semantic relation can be translated in an enormous quantity of textual patterns.

More limitations are related with the *Yahoo search engine*, which is not sufficiently versatile to make use of a large number of expressions. Besides that, by searching a flexed term, the search engine is not capable of search terms that have the same lemma, which limits this kind of search.

Still though, we moved on to the next step which consisted on verify if there was a correlation between the values obtained with the metrics for each type of relation and the human evaluation. However, due to the factors referred above and due to the low quantity of triples available to that calculation, we always obtained values with low correlation, that never exceeded 20%.

In the future, we pretend to continue our search for a automatic validation method to our work and we want to experiment these metrics in corpus to whom exist a more versatile search interface, like AC/DC (Costa et al. (2009)).

4.2.4 Conclusions

With this experiment, we have shown the performance of our system over the Wikipedia abstracts, which can be seen as semi-structured text, written in Portuguese. The knowledge extracted from it tends to be somewhat prototypical in nature, however in general should correspond well to the kind of information that lexicographers would expect in a lexical ontology.

Has we have seen by the outcomes, the system still has some problems that interfere with its performance. Some of these limitations are not only related with ambiguity, high complexity and the amount of ways to indicate a semantic relation of interest, but also related with the POS tagger used, and its lemmatiser. Whenever it does not recognize a word, it tries to infer its grammatical category based on probabilities, and its lemma based on rules. Sill, this process involves complex

tasks, such as understanding context and determining the part of speech of a word in a sentence (requiring, for example, knowledge of the language grammar), it can be a hard task to implement a lemmatiser for a Portuguese language. Thereat, currently, it becomes impossible to obtain triples whose arguments are lemmatised, because the lemmatiser used could deteriorate its quality.

Although we have found a way to filter almost every NER entries through DBpedia's taxonomy, there still exists various Wikipedia entries that can be interesting to analyse and they are being filtered unnecessary. So, we will continue to search for a better and most adequate method to filter these entities.

Besides some limitations on the process, we have proved that it is possible to validate the knowledge extracted from the Wikipedia abstracts in the *term-document* matrix, created from the CETEMPúblico corpus. For this particular experiment, the LSA metric gives more weigh correlation values than Cocitation, different from the *Experiment 1*. Although, we can not conclude that one metric is more appropriate than the other, leading us to induce that all metrics are important in different contexts.

Additionally, in the future we pretend to create a specific method to improve the precision of the hypernymy relations, obtained through the analysis of multi-words. Even though they are usually correct, the triples obtained with this method are normally too generic or obvious, which means they are useless in the enrichment or even in the creation of lexical ontologies. We believe this new method needs to take into account the number of occurrences (frequency) of the various atoms on the multi-word entities in the collections of documents, weighting these entities this way.

Furthermore, and despite all the efforts, the results obtained by the automatic evaluation method can only be considered theoretical. In the future, we pretend to continue our search for a better automatic evaluation method. Using a more versatile search interface, like the AC/DC interface (Costa et al. (2009)), studying and proposing other indicative patterns of interest.

Finally, with the results obtained in this experiment would be interesting to do an analysis in the quantity of extracted knowledge that is not included in other resource, like PAPEL (Gonçalo Oliveira et al. (2010b)), similar to what Hearst (Hearst (1998)) did for WordNet, see section 4.4.

4.3 Experiment 3: studying the system improvements

This section presents the experiment carried out to study the improvements of the system, described in section 3. To do that, the same system version used in *Experiment 2*, section 4.2, was used over the same textual resource used in the *Experiment 1*, section 4.1.

4.3.1 Experiment Goals

Like we have referred in the section 3, the main goal of our work is the creation of a system that automatically extracts knowledge from text, independent of their type (e.g. journalistic corpus, free text, etc.), and analyse the benefits of applying

metrics in this knowledge.

So, the main scope of this experiment is to study how the system has improved in the last experiment. To do that, we compare the percentage of triples from the *Experiment 1*, to the new triples extracted with the system presented in the *Experiment 2*, both manually evaluated.

Lastly, corpus distributional metrics presented in section 2.5.1 is used to study their application in the knowledge validation task, comparing its correlation coefficient values to *Experiment 1*.

4.3.2 Experiment

Experiment Set-up

Figure 4.13, presents all the modules used over the CETEMPúblico corpus[29].

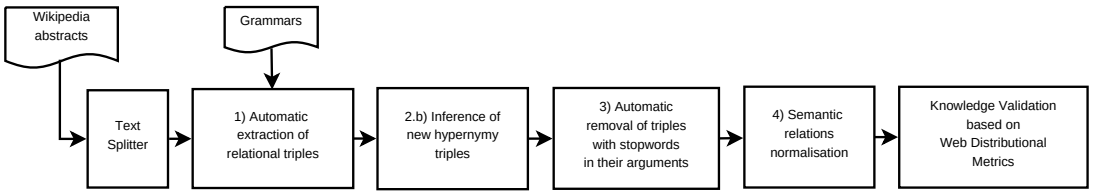


Figure 4.13: Modules used in *Experiment 3*.

This approach is based in Portuguese patterns, that are indicators of semantic relations in text, like is referred in the chapter 3. Some of them can be found in table 4.11, section 4.2.

The system developed has a modular architecture, where each module is independent of each other. Each one performing a specific task and have their own output file, making their maintenance and their use very versatile.

So, the system used in the *Experiment 2*, figure 4.13 could be easily changed to extract knowledge from the CETEMPúblico corpus, used in the *Experiment 1*. More specifically, we just have changed the **Text Splitter** module, in order to read the corpus text.

At last, like we have referred above, we have used the same documents used in *Experiment 1*. The first 28,000 documents of CETEMPúblico, which contain 30,100 unique content words (considering only nouns, verbs and adjectives) and results in approximately 1 million of *word-in-document* relations, called *term-document* matrix. To understand the concept *word-in-document* or simply *term-document* matrix, see the topic *Experiment Set-up* on section 4.1.

Experiment Results

Table 4.15 describes the quantity of triples extracted by relation, before (Total) and after removing (After Rem.) repeated triples, column *Experiment 3*.

Column *Experiment 1* presents the number of triples extracted after: removing repeated triples, **Automatic removal of triples with stopwords in their arguments** module, and performed the **2(a) Inference of new hypernymy triples**

[29]The CETEMPúblico version used in the experiment is non annotated. We will take advantage of our own POS tagger, the same adopted in the *Experiment 2*.

module analysis (After Rem. & Inf.), see *Extraction Results* topic in section 4.1 for more details.

	Experiment 3		***Experiment 1***
Relation	**Total**	**After Rem.**	**After Rem. & Inf.**
Hypernymy	306,762	286,960 (-6.5%)	9,365
Causation	5,691	3,037 (-46.6%)	2,660
Purpose	5,374	3,779 (-29.7%)	3,288
Part_of	2,316	1,759 (-24.1%)	1,373
Synonymy	284	254 (-10.6%)	270
TOTAL	*320,427*	*295,789 (-7.7%)*	*16,956*

Table 4.15: *Experiment 3* vs. *Experiment 1* - number of triples extracted from the CETEMPúblico corpus.

As we can see, the system have extracted more triples than in *Experiment 1*, specially for hypernymy relation, where is extracted ≈30.5% more triples. 281.944 triples were extracted by the **Automatic extraction of relational triples** module; 24.818 by the **2(b) Inference of new hypernymy triples** module. The resulted number of triples after perform after the **Automatic removal of triples with stopwords in their arguments** module is 269.385, ≈-4.5%, for **Automatic extraction of relational triples** and 17.575, ≈-29.2%, for **2(b) Inference of new hypernymy triples**.

The weighty number of triples extracted in this experiment can be explained by two reasons. The first one is related to the fact that the **2(b) Inference of new hypernymy triples** module used, is based in the 2(b) method, described in section 3.2. The extraction of new triples based on multi-word terms - 2(b) method - takes advantage of two lexical-syntactic patterns: [**N ADJ|ADV**] and [**N de|do|da|com|para N**]. Different from the adopted method in *Experiment 1*, that only infers new triples from complex entities on previously extracted triples, section 3.2. The second reason was related to the system improvements, more specifically in the hypernymy grammar, where we have eliminated some grammar rules, and added new ones.

However, we want to know if, with an higher number of triples extracted, the number of correct triples have increased or decreased. To do that, the precision of the manual evaluation results between the experiments, will be analysed.

Manual Evaluation

To evaluate the precision of the results, random samples for each type of relation, were selected. The samples' sizes took the type of relation into consideration and were divided into two random samples, each one evaluated by one human judge.

Each human judge was asked to assign one of the following values[30] to each triple, according to its quality:

- **0**, if the triple is completely incorrect.
- **1**, if the triple is not incorrect, but something is missing. Like a preposition or an adjective that makes one of the arguments strange and prevents the triple

[30]The used values in this experiment are the same used in the *Experiment 1*, allowing its comparison.

from being correct in one or both of its arguments, or even the relation is very generic.

- **2**, if the triple is correct.

The results obtained for manual evaluation are presented in the table 4.16. The total number of triples manual evaluated, and their percentage from the total extracted, are showed in column Evaluated and % from the Total, respectively.

	Manual Evaluation							
	0		1		2		**Evaluated**	**% from the Total**
Synonymy	40	43%	27	27%	25	27%	92	36%
Hypernymy	141	28%	165	33%	193	39%	499	0.2%
Causation	86	53%	55	34%	20	12%	161	5.3%
Purpose	101	52%	57	29%	37	19%	195	5.2%
Part_of	46	47%	38	39%	13	13%	97	5.5%

Table 4.16: Quantity of triples extracted and its manual evaluation results.

The percentage of values obtained for manual evaluation, in both experiments, are presented in the figure 4.14, where for the *Experiment 1* values, the name of the relation starts with a lowercase[31] letter, and for the results obtained in this experiment starts with an uppercase[32] letter.

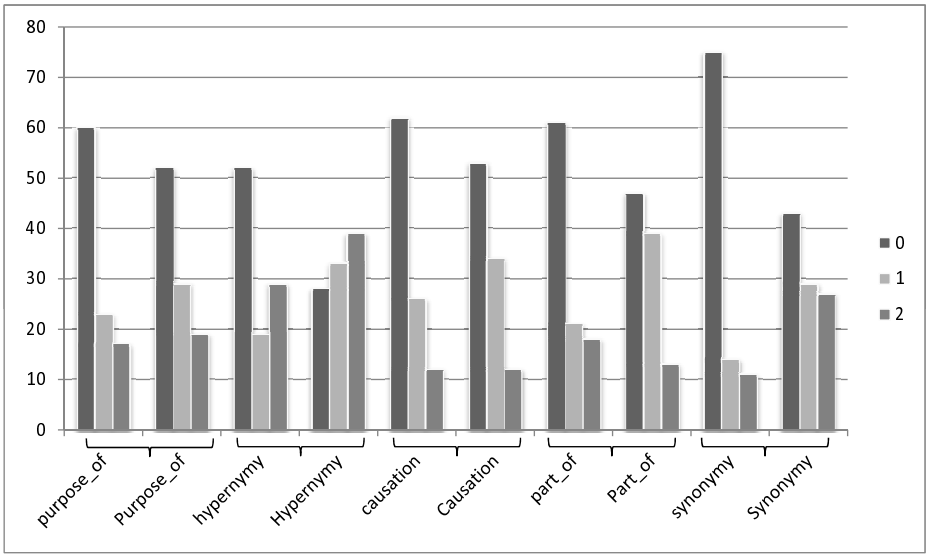


Figure 4.14: CETEMPúblico manual evaluation (first vs. second approach).

An interesting point, when we compare the manual results of both approaches, is the difference in the number of triples classified with 0, which decrease in all semantic relations. The greater pronounced decrement can be seen in the hypernymy relation.

Other relevant information is the number of correct triples. Its percentage increased in ≈10% more for hypernymy and synonymy relations. The better results,

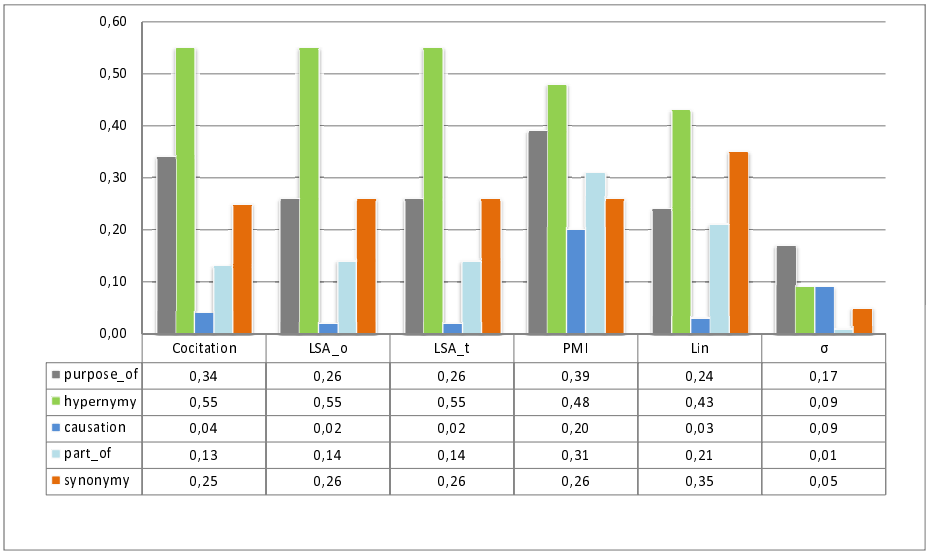
[31]I.e. purpose, hypernymy, causation part_of and synonymy.
[32]I.e. Purpose, Hypernymy, Causation Part_of and Synonymy.

for synonymy, can be explained by the fact that a lot of the Wikipedia's abstracts begins with enumerations. So, we have learned and consequently improved its semantic patterns in the grammar. The aforementioned explanation can also be applied to the hypernymy relation, because with the Wikipedia abstracts we have study the hypernymy patterns efficiency, and so, with *Experiment 2* we have eliminated some grammar rules that led to extract incorrect triples, and added new ones.

Nevertheless, triples classified with 1 are increased in all the relations, that essentially due to the POS tagger limitations. It does not recognise some words, leading to incorrect morpho-syntactic classification. Also due to the type of the text processed, most part of the triples which were changed in the final phase, **Semantic relations normalisation** module, associated to the fact that one or both of its arguments were not nouns, the triple normalisation was incorrect performed. However, this could be improved if we had a better POS tagger.

Manual Evaluation vs. Distributional Metrics

In order to observe the relationships between the manual evaluation and the output values given by the metrics, the correlation coefficients between them were computed and showed in figure 4.15. For this purpose, the correlation coefficient presented in equation 2.16, section 2.5.3, will be used in the same way that we have adopted in the last experiments.



	Cocitation	LSA_o	LSA_t	PMI	Lin	σ
purpose_of	0,34	0,26	0,26	0,39	0,24	0,17
hypernymy	0,55	0,55	0,55	0,48	0,43	0,09
causation	0,04	0,02	0,02	0,20	0,03	0,09
part_of	0,13	0,14	0,14	0,31	0,21	0,01
synonymy	0,25	0,26	0,26	0,26	0,35	0,05

Figure 4.15: Correlation coefficients (version 2) between manual evaluation and the corpus distributional metrics.

It is possible to observe that all the metrics have positive correlation values, a different scenario from the *Experiment 1*, where, for example to synonymy relation, it was not true, see figure 4.3. However, σ have low correlation coefficients in all the semantic relations. The reason can be because it is not based on co-occurrences, but on the relevance of the terms in text.

In the *Experiment 1*, we have concluded that synonymy entities, despite sharing very similar neighbourhoods, may not co-occur frequently in the same sentence

(Dorow (2006)) or even in the same document, because they are alternative names for the same thing. Nevertheless, the rules used to extract synonymy relations improved, resulting in a better correlation values, ≈28% more in average.

Similar correlation coefficients are obtained for the hypernymy relation, in both experiments, by all metrics, which suggests that hyponyms and their hypernyms tend to co-occur very frequently in text, like we have concluded in *Experiment 1*.

For purpose and causation, the results have improved, however can not be considered very significantly in the validation task. This happens because there are more ways to denote these relations in corpora, which led to less extracted and more incorrect triples. This is in conformity with the experiment described in Gonçalo Oliveira et al. (2009), where the patterns that denotes these relations were searched in the CETEMPúblico, in order to validate the PAPEL semantic triples.

On the other hand, part_of relations have worst results for Lin's measure and LSA; metrics that in *Experiment 1* have the better results, however in this experiment PMI seems to be the most correlated metric.

Another conclusion is that, with this experience, the obtained values for LSA calculated with the occurrences of the entities (LSA_o) are the same to the ones calculated with the TF-IDF (LSA_t). Same conclusion that in *Experiment 1*.

4.3.3 Conclusions

With this experiment, we have demonstrated the improvements of our system during the last two experiments, more precisely their impact in the precision of knowledge extracted from text. To do that, the same system version used in *Experiment 2*, section 4.2, has been used over the same text used in the *Experiment 1*, section 4.1.

Although this work is made for Portuguese, we believe that it can be adapted to other languages with similar distributional behaviour.

Most of the works similar to ours, but for English, propose using LSA-based filters. However, despite very close correlation results, in average, our adaptation of the Cocitation metric, seems to be the most adequate for such a task.

Besides high correlation values for hypernymy, the results for the others semantic relations resulted in low correlation. We thing this is understandable, because semantic terms can co-occur in many ways, and each semantic relation can be translated in an enormous quantity of textual patterns.

Inspired by Cederberg and Widdows (2003) and Wandmacher et al. (2007) work, we have computed LSA based on a TFI-IDF (LSA_t), although, in this experiment LSA calculated with the occurrences of the entities (LSA_o) have returned the same coefficient values - in *Experiment 1* we have concluded the same, so, we think the both approaches returns similar values.

4.4 Experiment 4: knowledge-bases comparison

A comparative view of the knowledge contained in three automatically created knowledge-bases is presented in this section. More specifically, we have compared the knowledge extracted by our system from CETEMPúblico, and from the Wikipedia abstracts, with the knowledge extracted from a Portuguese dictionary.

In this section, we start by presenting the experiment goals and set-up, and then,

the results obtained using the method presented in section 3.4.

In the end, some considerations are discussed in the section 4.4.3.

4.4.1 Experiment Goals

The extracted knowledge, already structured, can be useful to augment lexical resources. However, it would be interesting to analyse the quantity of new knowledge in each resource comparing to each other, similar to what Hearst (Hearst (1998)) did for WordNet.

So, in addition to our objective, which is to extract knowledge and validate them; we want to compare the knowledge extracted from CETEMPúblico and Wikipedia abstracts, *Experiment 3 and 2* respectively, with the knowledge in PAPEL; and how the three resources can be complementary.

Therefore, we have changed the method proposed by Hearst (Hearst (1998)), that only takes in consideration the hypernymy relation, and we have applied it to the five semantic relations covered by our system. Section 3.4, describes this method in detail.

4.4.2 Experiment

Experiment Set-up

For an easier experiment understanding, lets call CTPR to the knowledge extracted from the CETEMPúblico and, WIKIR to the knowledge extracted from the Wikipedia abstracts; more precisely the semantic knowledge result from the *Experiment 3* and *Experiment 2* (set **A2**), respectively.

To perform this experiment we have used the comparison method presented in section 3.4, more specifically, the algorithm 1, proposed by us. However, we needed to put the CTPR and WIKIR hypernymy triples, and the triples inferred in the same set, in order to standardize the knowledge-bases.

The third knowledge-base used in this experiment is the second version of PAPEL[33] (Gonçalo Oliveira et al. (2010b)), provided by Linguateca[34] (see *PAPEL* topic in section 2.4.2). This resource was downloaded and inserted in a database, that we will call PAPEL knowledge-base. It's triples are in the same formate used by us, $t_i = (e_1, r, e_2)$. However, in order to compare our knowledge resources with PAPEL, we have done some modification:

- we have duplicated the number of triples for synonymy, because in our system if, we extract the triple $t_i = (e_1, s, e_2)$ a new triple $t_j = (e_2, s, e_1)$ is created, where $\{_{i,j}\} \in \mathbb{N}$ and s is the initial letter for synonymy relation. For example, if PAPEL contains the triple *quickly* SYNONYM_OF *speedily* a new triple *speedily* SYNONYM_OF *quickly*, will be created.

- we have joined member with the part_of relation, because in our system we consider these two relation as one.

[33] http://linguateca.pt/PAPEL
[34] http://www.linguateca.pt

Table 4.17, presents an overview of the number of triples existing in the three knowledge-bases: PAPEL, WIKIR and CTPR. For the last two mentioned, we present the percentage of correct triples manually evaluated too, column % C.T.

Relation	PAPEL	WIKIR	% C.T.	CTPR	% C.T.
Synonymy	79,161	11,862	70%	254	27%
Hypernymy	62,591	45,791	34%	286,960	39%
Causation	7,966	520	22%	3,037	12%
Purpose	8,312	743	26%	3,779	19%
Part_of	13,389	1,287	23%	1,759	13%

Table 4.17: Total number of triples and their correctness percentage.

Not surprisingly, the biggest percentage of correct triples are associated the WIKIR. As we have concluded in the last experiments, the triples extracted from the journalistic corpus are more influenced by the context in with they appear, than those found in encyclopedia. Moreover, they tend to reflect subjective judgements, metaphorical usage, or opinions that the more established statements that appears in the encyclopedia.

Experiment Results

The number of triples and their percentage, by semantic relation, that PAPEL and CTPR have in common, is presented on table 4.18, where PAPEL correspond to the DB_b and CTPR to DB_a. As we can see, there is no common knowledge between the two resources, except for hypernymy that shares 15 triples. Besides that, there are none triples that have both entities on PAPEL, related with a different relation. This is not surprising, still they were created from two distinct types of text; PAPEL from a dictionary and CTPR from a journalistic corpus.

Table 4.19 presents the common triples and their percentages in both knowledge-bases, with its number of triples, where PAPEL correspond to the DB_b and WIKIR to DB_a. Not surprising, WIKIR shares more triples then CTPR in PAPEL, essentially due to the fact that WIKIR was created from a general resource, an encyclopidia, that contains knowledge related to plants, animals, etc. - some of them included in a dictionary. An interesting observation is the number of triples in *C2*, that in this particular case, all of then should belong to the set *C1*, however due to the system POS tagger limitations, the entities were wrongly classified (e.g. in WIKIR the entities *convertível* and *conversível* were classified as nouns resulting in *convertível* SINONIMO_N_DE *conversível*, however in PAPEL they were classified as adjectives: *convertível* SINONIMO_ADJ_DE *conversível*).

Having this in mind, it is normal that CTPR and WIKIR almost do not share knowledge, see table 4.20, where WIKIR correspond to the DB_b and CTPR to DB_a. Once again, the corpus used to create this two knowledge-bases are distinct, one covers journalistic events and the other general knowledge.

4.4.3 Conclusions

Whereas PAPEL was created from a Portuguese dictionary, containing restrict knowledge about the words and their meanings, WIKIR was created by an ecyclopedia. It contains not only knowledge related to words, but more about knowledge

associated to the world and human knowledge. For this reason, it shares more knowledge then CTPR, that was created by a journalistic corpus, which covers specific knowledge.

Additionally, we can conclude that these three resources are complementary between them, and they can be joined to define more knowledge about the World, and creating, that way a broad-coverage resource.

Nevertheless, in this experiment a method that compares different knowledge-bases, was presented and, we think this can be useful in the ontology creation task. More precisely, this method can be used to study the common knowledge between two knowledge-bases, and their completeness.

	Number of Triples		C1		C2		C3		C4		C5	
	PAPEL	CETEMPúblico	N	(≈)	N	(≈)	N	(≈)	N	(≈)	N	(≈)
Synonymy	79,161	254	2	0.8%	0	0%	80	31.5%	76	29.9%	96	37.8%
Hypernymy	62,591	286,960	15	0%	0	0%	1,010	0.4%	199,115	69.4%	86,820	30.2%
Causation	7,966	3,037	0	0%	0	0%	681	22.4%	1,532	50.5%	824	27.1%
Purpose	8,312	3,779	0	0%	0	0%	1,000	26.5%	1,879	49.7%	900	23.8%
Part_of	13,389	1,759	3	0.9%	0	0%	277	15.7%	874	49.7%	605	33.7%

Table 4.18: Common knowledge between CTPR and PAPEL knowledge-base.

	Number of Triples		C1		C2		C3		C4		C5	
	PAPEL	Wikipedia	N	(≈)	N	(≈)	N	(≈)	N	(≈)	N	(≈)
Synonymy	79,161	11,862	115	1%	7	0.1%	310	2.6%	3,220	27.1%	8,210	69.2%
Hypernymy	62,591	45,791	137	0.3%	0	0%	496	1.1%	28,011	61.2%	17,147	37.4%
Causation	7,966	520	0	0%	1	0.2%	61	11.7%	218	41.9%	240	46.2%
Purpose	8,312	743	0	0%	0	0%	33	4.4%	279	37.6%	431	58%
Part_of	13,389	1,287	2	0.2%	0	0%	19	1.5%	261	20.3%	1,005	78%

Table 4.19: Common knowledge between WIKIR and PAPEL knowledge-base.

	Number of Triples		C1		C2		C3		C4		C5	
	Wikipedia	CETEMPúblico	N	(≈)	N	(≈)	N	(≈)	N	(≈)	N	(≈)
Synonymy	11,862	254	0	0%	0	0%	6	2.4%	44	17.3%	204	80.3%
Hypernymy	45,791	286,960	3	0%	0	0%	995	0.3%	91,040	31.7%	194,922	68%
Causation	520	3,037	1	0%	0	0%	30	1%	591	19.5%	2,415	79.5%
Purpose	743	3,379	0	0%	0	0%	59	1.7%	807	23.9%	2,513	74.4%
Part_of	1,287	1,759	0	0%	0	0%	10	0.6%	280	15.9%	1,469	83.5%

Table 4.20: Common knowledge between CTPR and WIKIR knowledge-base.

Chapter 5

Conclusions and Future Work

To realise the semantic knowledge representation vision of a world, where the "world meaning" is not always conceptual, mechanisms must be developed to represent and reason the uncertainty that knowledge can arise. In this thesis, extraction methods developed to the English language are adapted, with some new methodologies in order to unify, learn and reason about world knowledge representation. Taking in consideration the distributional hypothesis (Harris (1970)), which assumes that similar words tend to occur in similar contexts, we have computed the distributional metrics between words, to study and take conclusions of their application in the improvements of the knowledge extracted by our system. The main contributions of this thesis are:

- If, on the one hand, the modules that extract written data from textual resources were never been used in our experiments, on the other hand, they provide various types of libraries, that are capable of interpreting text contained in different textual sources, such as docs, pdfs, etc., and even it is capable of extract text from Web pages.

- Unlike other approaches, we studied not only about the most used lexico-syntactic patterns that are presented in the literature, but we have also proposed new indicative patterns to the five semantic relations covered by our system.

- We have studied and proposed a method to infer the hypernymy relation from multi-words and also from triples already extracted by our system.

- We have shown that the precision of systems capable of acquiring semantic knowledge from text may benefit from applying distributional metrics to their output. Although this work is made for Portuguese, we believe that it can be adapted to other languages with similar distributional behaviour.

- We proposed a new method, based on Web metrics and lexico-syntactic patterns, to automatic evaluate semantic knowledge extracted from text.

- We developed a method to compare knowledge-bases. This method is very simple, but at same time give us an important overview about the common knowledge between them. It is independent from the language, and can be applied to several domains, where there is no other manner to know the quantity of knowledge shared between knowledge-bases.

- Our system can be used in many crucial areas for NLP task, such as information extraction (IE) (from the most popular formats), lexical ontology creation and information quantification.

In fact, since our system is a so broad approach, we think that our approach can be seen as an introductory step in the Portuguese semantic knowledge extraction task, resulting in several modules that can be used in a broad ontology creation for the Portuguese language. Also, it performs the first attempt to validate knowledge on the Portuguese language, with interesting results for some semantic relations.

However, currently, the system contains some problems that interfere with its use in any kind of text. The high difficulty in defining textual patterns able to extract all instances of a particular relation of interest, with the possibility of transmitting the same idea in different ways, increases the ambiguity of text; and also the non existence of boundaries on the vocabulary used, with the existence of anaphora; largely restricts its broad appliance in all domains.

This thesis has achieved several deliverables, completed at different stages of the project. At the end of the project, the following milestones were delivered:

- **Bibliographic Revision:** describing the most relevant concepts to this thesis, chapter 2;

- **Resources and Tools Study:** presenting the tools, libraries and resources that could be used in this thesis, section 2.6 and 2.4;

- **Similarity Distributional Metrics:** describing statistical methods, usually used in information retrieval (IR) tasks, section 2.5;

- **Thesis Proposal Elaboration:** describing the proposed approach and all similar approaches to ours;

- **System Architecture:** presents in detail all the system modules and their purpose, chapter 3;

- **First Experiment:** an experimental approach in CETEMPúblico, a journalistic corpus, using a simple version of the system, section 4.1;

- **Second Experiment:** an approach in the Wikipedia abstracts, which consist of the first sentences in encyclopedia articles, using a better version of the system, section 4.2;

- **Third Experiment:** a second approach in the CETEMPúblico corpus, comparing the first and the second version of the system, section 4.3;

- **Fourth Experiment:** presents an experimental approach (using the proposed method in section 3.4), that analyses the quantity of common knowledge between three resources, one automatically extracted from a dictionary and the other two obtained with our experimentation, section 4.4;

- **Final Thesis Elaboration:** describing all the work done in this thesis.

The work plan defined for this thesis, with tasks and respective schedule, is represented in figure 5.1. Nevertheless, some deliverables suffered some delays compared to the schedule, but they were completely finished as shows figure 5.2. Furthermore, additional tasks were added to the initial schedules, because we believe they were the best directions to follow, in order to complete this research.

5.1 Publications

The experiment performed in the CETEMPúblico corpus (*Experiment 1*), originated one publication in the 19^{th} European Conference on Artificial Intelligence (ECAI 2010)[1], more specifically in the Language Technology for Cultural Heritage, Social Sciences, and Humanities workshop (LaTeCH 2010)[2], namely Costa et al. (2010) (available in the next url: `http://student.dei.uc.pt/~hpcosta/papers/ecai2010.pdf`).

The work presented in Costa et al. (2010) analyses the benefits of applying metrics based on the occurrence of words in documents to a set of relational triples automatically extracted from corpora. This experiment uses a simple system to extract triples automatically from a corpus. Then, the same corpus is used for weighting each triple according to well-known distributional metrics. Finally, some conclusions are presented on the correlation between the values given by the metrics and the evaluation made by humans.

Besides the latter publication, a paper describing the performance of our system in the Wikipedia abstracts (*Experiment 2*), was written and accepted in the 2^{nd} Informatics Symposium (INForum 2010)[3], more specifically in the management and treatment of information track, namely Gonçalo Oliveira et al. (2010a) (available in the next url: `http://student.dei.uc.pt/~hpcosta/papers/inforum2010.pdf`).

The aforementioned paper describes the system that was applied to Wikipedia, currently a huge and free source of knowledge. The obtained results are shown and their evaluation is discussed together with the current limitations and cues for further improvement.

ECAI is the leading Conference on Artificial Intelligence in Europe. Since it is an international conference we have tried to make a position in the LaTeCH workshop, presenting our work. INForum is a national event that intends to put together researchers and professionals in the areas of Databases, Data Mining, IR and NLP, which scientific work can be complementary to solve current problems in areas related to information managing and processing.

5.2 Future Work

In this section, we identify some directions to future work. Some of these directions have the purpose of improving the executed work, while others explore some new interesting ideas and concepts that aroused us during this work.

[1] http://ecai2010.appia.pt
[2] http://ilk.uvt.nl/LaTeCH2010
[3] http://inforum.org.pt/INForum2010

5.2.1 General Ideas

In general, there are several issues that deserve further exploration. The most obvious is the system application in more different domains. We have applied our system in several relevant types of text, such as CETEMPúblico, a journalistic corpus (section 4.1) and Wikipedia, an encyclopidia (section 4.2), however more domains, like restrict corpus or even free text (documents, Web pages, etc.) could be used to improve our system, more specifically the **Knowledge Extraction** modules.

Another important improvement would be the **Knowledge Validation**. Although we have explored several ways to automatically validate semantic knowledge extracted from text, there is still much to do.

Also, more ideas could be explored:

- **Discovery on new semantic patterns:** we have explored the use of one textual resource to learn and infer new lexico-syntactic patterns, that indicate same relation of interest. Other way to do that would be though machine learning techniques. Using a set of correct and incorrect entities in a bigger corpus, such as the Web.

- **Extraction of semantic knowledge from text:** Despite having used PEN as a tool to extract semantic knowledge from text, we consider it to be limited. The use of grammars, created by human enforce, is a limitation, because they can not predict all the lexico-syntactic patterns that indicate some relation of interest. So, the use of machine learning techniques would improve substantially our approach, besides that would be more broad-coverage and versatile as regards the variations in lexico-syntactic patterns, contained in the middle of two entities.

- **Studying the better windows size:** Although our *term-document* matrix experimentation dominates *term-term* matrix in 89%, further studies are needed to understand how this influences the corpus distributional metrics values and their application to a system like ours. Also, more experimentation would be needed to select the best window size.

- **Broad coverage database:** Even though we obtained interesting results using distributional metrics in the *term-document* matrix, sometimes, the returned values did not reflect the expected probability, giving inconsistent probabilities to some relations of interest. It is studied (Dorow (2006)), that some words, despite sharing very similar neighbourhoods, may not co-occur frequently in the same sentence. However, if we have a bigger matrix and created from more different kinds of text, we believe that can be mitigated.

- **Weighting triples:** The idea is to study how other resources can be used, like TeP and PAPEL, to verify if the extracted triple $t_i = e_1, r, e_2$, or its entities, are already present in one, or in both resources, in order to assign weights to these triples. Other idea, is to weighting the entities based on their occurrence in some textual resource.

- **Evaluation module:** Besides the automatic evaluation proposal (section 4.2.3), that can be considered as the first step, it would be interesting a deeper

study, in order to understand its possible use in an additional module in our system, or even in others IE systems.

5.2.2 System

One of the most interesting future tasks would be to develop a complete Java framework, that could be easily used to extracts semantic knowledge from any kind of text, e.g. corpus, Web, documents, etc. This framework would be especially developed for IE tasks, more specifically, optimised for extraction and quantification of lexico-semantic knowledge from text. It would use the most adequate IE algorithms, depending the kind of text in use, as well its IR methods. This system would also help and provide their modules, that may be broadly used by researchers and developers that work with Portuguese NLP.

Knowledge Extraction

There are many ways that the structure of a language can indicate the meanings of lexical items, however to NLP tools the main difficulty lies in finding constructions that frequently and reliably indicate a relation of interest. In the future, we want to reduce this ambiguity using more lexico-syntatic patterns. Nevertheless, with our experience we are aware that is not enough, so new automatic methods to validate and evaluate the knowledge extracted with our system need to be improved or even created.

Validate and Evaluate Knowledge

In our thesis, we have shown that the precision of our system would benefit from applying metrics, normally used in IR task. Nevertheless, we have proved that one metric is not better than the others in all relations of interest. Besides that, the use of machine learning techniques, that combine several metrics minimises the square error. We believe that is the way to automatic validate knowledge extracted from text.

An additional step, is the use of techniques to measure the semantic similarity between entities. To do that, we have explored the redundancy and size of a huge corpus, the World Wide Web. However, due to the search engines limitation, and also the difficulty in create semantic patterns that (almost) always indicate the semantic relation of interest, the outcomes can only be considered theoretic. Nevertheless, in the future we will spend more time to investigate this approach: combining Web metrics with lexical patterns, using machine learn techniques, or even use a static corpus that allow us, the use of a more versatile queries, like the interface AC/DC (Costa et al. (2009)).

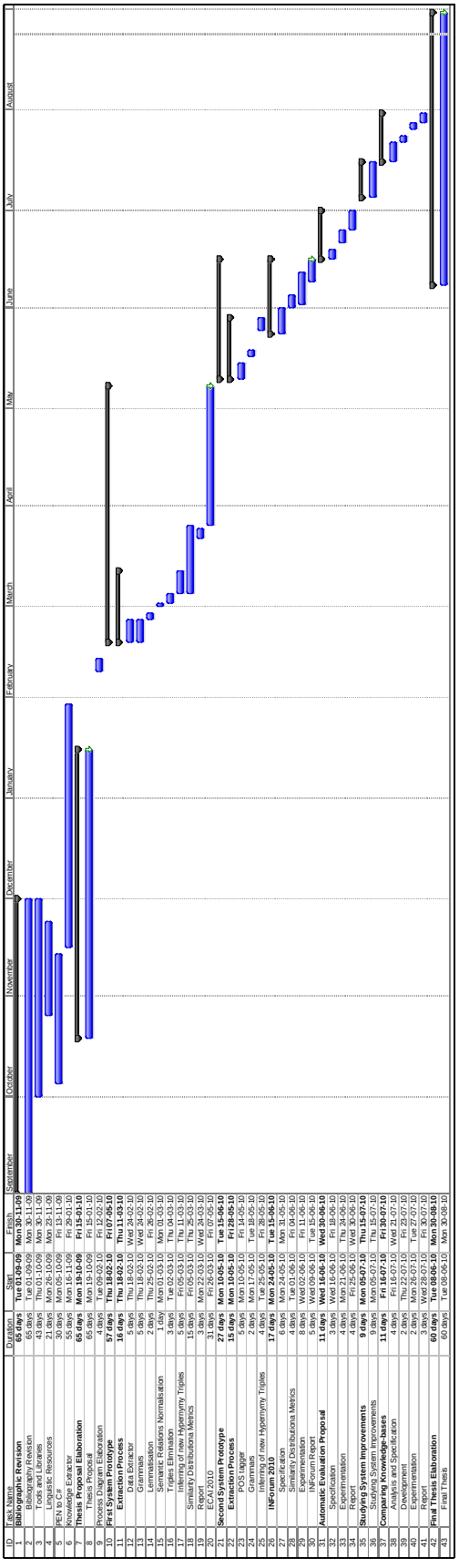


Figure 5.1: Project planning proposal.

Figure 5.2: Final project planning.

References

Agichtein, E. and Gravano, L. (2000). Snowball: Extracting Relations from Large Plain-Text Collections. In *Proc. 5th ACM International Conference on Digital Libraries*, pages 85–94.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., and Padró, M. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proc. 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 48–55.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proc. 17th International Conference on Computational Linguistics*, pages 86–90, Morristown, NJ, USA. ACL.

Barlow, R. E., Bartholomew, D., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions: the theory and application of isotonic regression*. Wiley, New York.

Baségio, T. L. (2007). *Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul PUCRS.

Bennet, C. H., Gcs, P., Li, M., Vitanyi, P. M. B., and Zurek, W. (1998). Information Distance. In *IEEE Trans. Information Theory*, volume 44, pages 1407–1423.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 57–64, Morristown, NJ, USA. ACL.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web: Scientific american. *Scientific American*.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Protuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University, Arhus.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

Blohm, S., Cimiano, P., and Stemle, E. (2007). Harvesting relations from the web: quantifiying the impact of filtering functions. In *Proc. 22nd National Conference on Artificial Intelligence (AAAI'07)*, pages 1316–1321. AAAI Press.

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In *Proc. 16th International Conference on World Wide Web*, pages 757–766, New York, NY, USA. ACM Press.

Boss, A. H. and Ritter, J. B. (1993). *Electronic Data Interchange Agreements - A Guide and Sourcebook*. Publication CCI No. 517, Paris.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proc. Biennial GSCL Conference 2009, Meaning: Processing Texts Automatically*, pages 31–40, Tbingen, Gunter Narr Verlag.

Budiu, R., Royer, C., and Pirolli, P. (2007). Modeling Information Scent: A Comparison of LSA, PMI and GLSA Similarity Measures on Common Tests and Corpora. In *Proc. RIAO*.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 120–126, Morristown, NJ, USA. ACL.

Cederberg, S. and Widdows, D. (2003). Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Proceedings of CoNLL*, pages 111–118.

Chinchor, N. and Robinson, P. (1997). MUC-7 named entity task definition. In *Proc. 7th Message Understanding Conference (MUC-7)*.

Chodorow, M. S., Byrd, R. J., and Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proc. 23rd annual meeting on Association for Computational Linguistics (ACL'85)*, pages 299–304, Morristown, NJ, USA. ACL.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proc. 27th Annual Meeting on Association for Computational Linguistics (ACL'89)*, pages 76–83, Morristown, NJ, USA. ACL.

Cilibrasi, R. L. and Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.

Cilibrasi, R. L. and Vitanyi, P. M. B. (2009). Normalized Web Distance and Word Similarity. *ArXiv e-prints*.

Cimiano, P. and Wenderoth, J. (2007). Automatic Acquisition of Ranked Qualia Structures from the Web. In *Proc. 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 888–895, Prague, Czech Republic. ACL.

Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2010). The Impact of Distributional Metrics in the Quality of Relational Triples. In *Proc. ECAI 2010, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH'10)*.

Costa, L., Santos, D., and Rocha, P. A. (2009). Estudando o português tal como é usado: o serviço AC/DC. In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL'09)*.

Costa, R. and Seco, N. (2008). Hyponymy Extraction Using the Pattern NP PREP NP in Portuguese Search Logs. Report, CISUC, Universidade de Coimbra, Portugal.

Cristo, M., de Moura, E. S., and Ziviani, N. (2003). Link information as a similarity measure in web classification. In *Proc. 10th Symposium On String Processing and Information Retrieval*, pages 43–55. Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

Dias-da-Silva, B. (2006). Wordnet.Br: an exercise of human language technology research. In *Proc. 3rd International WordNet Conference (GWC'06)*, pages 22–26, Jeju Island, Korea.

Dias-da-Silva, B. and Moraes, H. R. (2003). A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115.

Dias-da-Silva, B., Oliveira, M., and Moraes, H. (2002). Groundwork for the Development of the Brazilian Portuguese Wordnet. In *Proc. 3rd International Conference on Advances in Natural Language Processing (PorTAL'02)*, pages 189–196, London, UK. Springer.

Dorow, B. (2006). *A Graph Model for Words and their Meanings*. PhD thesis, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Earley, J. (1983). An efficient context-free parsing algorithm. *Commun. ACM*, 26(1):57–61.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in KnowItAll: (preliminary results). In *Proc. 13th International Conference on World Wide Web (WWW'04)*, pages 100–110, New York, NY, USA. ACM.

Fallaw, W. C. (1979). A test of the Simpson coefficient and other binary coefficients of faunal similarity. *Journal of Paleontology*, 53(4):1029–1034.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Fillmore, C. J. (1982). Frame Semantics. In of Korea, L. S., editor, *Linguistics in the morning calm*. Seoul: Hanshin Publishing Co.

Frankenberg-Garcia, A. and Santos, D. (2002). COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, IX(1):61–79.

Freitas, M. C. (2007). *Elaboração automática de ontologias de domínio: discussão e resultados*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.

Gaines, B. R. and Shaw, M. L. (1997). Knowledge acquisition, modeling and inference through the world wide web. *International Journal of Human-Computer Studies*, 46:729–759.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic Discovery of Part-Whole Relations. *Computational Linguistics*, 32(1):83–135.

Girju, R. and Moldovan, D. (2002). Text Mining for Causal Relations. In *Proceedings of FLAIRS Conference*, pages 360–364.

Girju, R., Putcha, M., and Moldovan, D. (2003). Discovery of Manner Relations and their Applicability to Question Answering. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Workshop on Multilingual Summarization and Question Answering*, pages 54–60.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27:153–198.

Gonçalo Oliveira, H. (2009). *Ontology Learning for Portuguese*. PhD thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering.

Gonçalo Oliveira, H., Costa, H., and Gomes, P. (2010a). Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In *Proc. INFORUM 2010 Workshop on Gestão e Tratamento de Informação (INFORUM'10)*.

Gonçalo Oliveira, H. and Gomes, P. (2008). Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário. Technical report, Linguateca.

Gonçalo Oliveira, H. and Gomes, P. (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proc. 5th European Starting AI Researcher Symposium (STAIRS'10)*, pages 1135–1136. IOS Press.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2009). Relations extracted from a Portuguese dictionary: results and first evaluation. In *Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA'09)*.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2010b). Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. 2(1):77–93. Nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009.

Gonçalo Oliveira, H., Santos, D., Gomes, P., and Seco, N. (2008). PAPEL: A Dictionary-Based Lexical Ontology for Portuguese. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Proceedings of Computational Processing of the Portuguese Language (PROPOR)*, volume 5190 of *LNAI*, pages 31–40. Springer.

Grishman, R. (1997). Information Extraction: Techniques and Challenges. In *International Summer School on Information Extraction (SCIE)*, pages 10–27.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Guarino, N. (1998). Formal Ontology and Information Systems. In *Proc. 1st International Conference on Formal Ontologies in Information Systems (FOIS'98))*, pages 3–15. IOS Press.

Guarino, N. and Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32.

Guthrie, L., Slator, B. M., Wilks, Y., and Bruce, R. (1990). Is there content in empty heads? In *Proc. 13th Conference on Computational Linguistics*, pages 138–143, Morristown, NJ, USA. ACL.

Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th Conference on Computational Linguistics*, pages 539–545, Morristown, NJ, USA. ACL.

Hearst, M. A. (1998). Automated Discovery of WordNet Relations. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.

Herbelot, A. and Copestake, A. (2006). Acquiring Ontological Relationships from Wikipedia Using RMRS. In *Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*.

Hirst, G. (2004). Ontology and the Lexicon. In *Handbook on Ontologies*, pages 209–230.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24:2–40.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Jackson, P. and Moulinier, I. (2002). *Natural Language Processing for Online Applications. Text retrieval, extraction and categorization*, volume 5 of *Natural Language Processing*. Benjamins, Amsterdam, Philadelphia.

Junior, L. C. R. (2008). *OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa*. PhD thesis, Centro de Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos, São Leopoldo.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall.

Khoo, C. S. G., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proc. 38th Annual Meeting on Association for Computational Linguistics (ACL'00)*, pages 336–343, Morristown, NJ, USA. ACL.

Knowles, J., Thiele, L., and Zitzler, E. (2006). A Tutorial on the Performance Assessment of Stochastic Multiobjective Optimizers. TIK Report 214, Computer Engineering and Networks Laboratory (TIK), ETH Zurich.

Kozima, H. and Furugori, T. (1993). Similarity between words computed by spreading activation on an English dictionary. In *Proc. 6th Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 232–239, Morristown, NJ, USA. ACL.

Lenat, D. (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38:33–38.

Lin, D. (1998a). An Information-Theoretic Definition of Similarity. In *Proc. 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

Lin, D. (1998b). Automatic retrieval and clustering of similar words. In *Proc. 17th International Conference on Computational Linguistics*, pages 768–774, Morristown, NJ, USA. ACL.

Liu, H. and Singh, P. (2004a). Commonsense reasoning in and over natural language. In *Proc. 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES-2004)*. Springer.

Liu, H. and Singh, P. (2004b). ConceptNet: A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22(4):211–226.

Marrafa, P. (2002). Portuguese WordNet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., and Mendes, S. (2006). WordNet.PT new directions. In Sojka, P., Choi, K., Fellbaum, C., and Vossen, P., editors, *Proc. 3rd International WordNet Conference (GWC'06)*, pages 319–320.

Matveeva, I., Levow, G., Farahat, A., and Royer, C. (2005). Terms representation with Generalized Latent Semantic Analysis. In *Proc. RANLP*.

Maziero, E. G., Pardo, T. A. S., Felippo, A. D., and Dias-da-Silva, B. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.

Michiels, A., Mullenders, J., and Noël, J. (1980). Exploiting a large data base by Longman. In *Proc. 8th Conference on Computational Linguistics*, pages 374–382, Morristown, NJ, USA. ACL.

Mineiro, A., Dória, M., Antunes, M., and Correia, M. (2004). Hiponímia e meronímia num corpus da naútica em português europeu. In *Actas do IX Simpósio Ibero-americano de Terminologia La terminologia en el siglo XXI, contribución a la cultura de la paz, la diversidad y la sostenibilidad.*, pages 361–380, Barcelona. IULA-UPF.

Morin, E. and Jacquemin, C. (2004). Automatic acquisition and expansion of hypernym links. *Computer and the Humanities*, 38(4):343–362.

Nichols, E., Bond, F., and Flickinger, D. (2005). Robust ontology acquisition from machine-readable dictionaries. In *Proc. 19th International Joint Conference on Artificial Intelligence(IJCAI'05)*, pages 1111–1116, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Technical report, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

O'Hara, T. P. (2005). *Empirical acquisition of conceptual distinctions via dictionary definitions*. PhD thesis, Las Cruces, NM, USA. Chair-Wiebe, Janyce.

Oliveira, P. C. (2009). Probabilistic Reasoning in the Semantic Web using Markov Logic. Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering.

Olney, J., Revard, C., and Ziff, P. (1967). Summary of some computational aids for obtaining a formal semantic description of English. In *Proc. 1967 Conference on Computational Linguistics*, pages 1–5, Morristown, NJ, USA. ACL.

Orengo, V. M. and Huyck, C. (2001). A Stemming Algorithm for Portuguese Language. In *Proc. 8th Symposium on String Processing and Information Retrieval (SPIRE 2001) - Chile*, pages 186–193.

Pantel, P. and Lin, D. (2002). *Discovering word senses from text*, pages 613–619. ACM, New York, NY, USA.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Resnik, P. (1999). Semantic Similarity in a Taxonomy An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130.

Richardson, S. D., Dolan, W. B., and Vanderwende, L. (1998). MindNet: acquiring and structuring semantic information from text. In *Proc. 17th International Conference on Computational Linguistics*, pages 1098–1102, Morristown, NJ, USA. ACL.

Riloff, E. and Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proc. 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 117–124.

Roark, B. and Charniak, E. (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proc. 17th International Conference on Computational Linguistics*, pages 1110–1116, Morristown, NJ, USA. ACL.

Rosenfeld, R. (1996). A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech and Language*, 10:187–228.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5):513–523.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for portuguese. In *Proc. 39th Annual Meeting on Association for Computational Linguistics (ACL'01)*, pages 450–457, Morristown, NJ, USA. ACL.

Santos, D. and Sarmento, L. (2003). O projecto AC/DC: acesso a corpora/disponibilização de corpora. In Mendes, A. and Freitas, T., editors, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL'02)*, pages 705–717, Lisboa. APL.

Schlenoff, C., Gruninger, M., Ciocoiu, M., and Lee, J. (1999). The essence of the process specification language. *Transactions of the Society for Computer Simulation*, 16(4):204–216.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. International Conference on New Methods in Language Processing*, pages 44–49.

Schreiber, G., Wielinga, B., de Hoog, R., Akkermans, H., and vd Velde, W. (1994). CommonKADS: A Comprehensive Methodology for KBS Development. *IEEE Intelligent Systems*, 9:28–37.

Simões, A. and Almeida, J. J. (2002). jspell.pm - um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas do XVII Encontro da Associação Portuguesa de Linguística*, pages 485–495, Lisboa.

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science*, 241(1):1–31.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open Mind Common Sense: Knowledge Acquisition from the General Public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 1223–1237, London, UK. Springer.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269.

Smullyan, R. M. (1995). *First-Order Logic*. Dover, NY.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning Syntactic Patterns for Automatic Hypernym Discovery. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, pages 1297–1304. MIT Press, Cambridge, MA.

Strzalkowski, T. and Harabagiu, S. (2006). *Advances in Open Domain Question Answering (Text, Speech and Language Technology)*. Springer, Secaucus, NJ, USA.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI–IR versus LSA on TOEFL. In Raedt, L. D. and Flach, P., editors, *Proc. 12th European Conference on Machine Learning (ECML'01)*, volume 2167, pages 491–502. Springer.

Uschold, M. and Grüninger, M. (1996). Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155.

Vanderwende, L., Kacmarcik, G., Suzuki, H., and Menezes, A. (2005). MindNet: an automatically-created lexical resource. In *Proc. HLT/EMNLP on Interactive Demonstrations*, pages 8–9, Morristown, NJ, USA. ACL.

Vossen, P. and Letteren, C. C. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proc. of the DELOS Workshop on Cross-language Information Retrieval*, pages 5–7.

Wandmacher, T., Ovchinnikova, E., Krumnack, U., and Dittmann, H. (2007). Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In Meyer, T. and Nayak, A. C., editors, *Third Australasian Ontology Workshop (AOW 2007)*, volume 85 of *CRPIT*, pages 61–69, Gold Coast, Australia. ACS.

Welty, C. and Guarino, N. (2001). Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering*, 39(1):51–74.

Witten, I. H. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Zero, S., Ivens, S., Millard, M., and Duvvuri, R. (1995). The Educator's Word Frequency Guide. In *Proc. Touchstone Applied Science Associates*.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and Grunert da Fonseca, V. (2003). Performance Assessment of Multiobjective Optimizers: An Analysis and Review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132.