Improving Semantic Similarity Measures for Word Pair Comparison

RAUL ERNESTO MENENDEZ MORA

DOCTOR OF PHILOSOPHY

Department of Informatics
School of Multidisciplinary Sciences
The Graduate University of Advanced Studies

2011

Improving Semantic Similarity Measures for Word Pair Comparison

by

RAUL ERNESTO MENENDEZ MORA

DOCTOR OF PHILOSOPHY

Supervisor: Ryutaro Ichise

Department of Informatics
School of Multidisciplinary Sciences
The Graduate University of Advanced Studies

Submitted
March, 2012

A dissertation submitted to the Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University of Advanced Studies (SOKENDAI)
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Advisory Committee

Assoc. Prof. Ryutaro Ichise	National Institute of Informatics, SOKENDAI
Prof. Akiko Aizawa	National Institute of Informatics, SOKENDAI
Prof. Hideaki Takeda	National Institute of Informatics, SOKENDAI
Prof. Seiji Yamada	National Institute of Informatics, SOKENDAI
Prof. Takahira Yamaguchi	Keio University

Abstract

The semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. In order to achieve the goals of the semantic web, it have to be able to define and to describe the relations among data (i.e., resources) on the Web. Ontologies are one of the formal representation for organizing information in the semantic web and they are also used in artificial intelligence, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it. In the semantic web context, since many actors provide their own ontologies, ontology matching or ontology alignment has taken a critical role for helping heterogeneous resources to inter-operate [23].

Ontology matching tools find classes of data that are "semantically equivalent". This process determine correspondences between concepts which are called alignments [22]. Finding those correspondences imply a semantic similarity assessment between the involved concepts.

Semantic similarity of words pairs is often represented by the similarity between the concepts associated with the words. Several methods have been developed to compute words similarity, most of them operating on taxonomic dictionaries like WordNet [24] or external corpus like the Brown Corpus. However the majority of them suffer from a serious limitation. They only focus on the semantic information shared by those words, or in the semantic differences, but they have been rarely combined in a broader perspective.

In this thesis we developed and applied a model of semantic similarity computation for word pair comparison. This model consider the semantic commonalities and the semantic differences as the core of its approach. By applying the model five new WordNet-based semantic similarity measures for word pair comparison were created. Four of this semantic similarity measures obtained higher values of correlation with human judgment than their original expressions, while the fifth one remained as competitive as their original version.

We also study WordNet taxonomic properties to extend a corpus-independent information content metric. The application of this new metric in one of the previously developed node-based semantic similarity allowed us to obtain the highest value of correlation with respect to human judgment. This thesis provides a general an extensible approach of semantic similarity computation for word pair comparison.

Dedication

To my family.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Ryutaro Ichise, for accepting me in his laboratory and for guiding me during all of these years that I have been studying at The Graduate University for Advanced Studies (SOKENDAI). He has been a priceless source of advices and support, always pushing me to give my best.

I thank SOKENDAI professors Hideaki Takeda, Seiji Yamada, Akiko Aizawa and Nigel Collier as well as professor Takahira Yamaguchi, for their time and valuable insights that have guided me during work on my thesis.

My eternal gratitude goes to my family for their constant support and encouragement, especially to my mother, my father (although not physically with us) and my brother. To my beloved wife and two kids that despite the physical distance, always gave me their positivism, trust and love. They make me feel so strong and they always were there for me in good or bad. To my friends: Sharkito, Danilo, José Alain, Orlando, Fito and Peligro which always help me out no matter what. Impossible not to mention here: uncle Camilo, my cousins José Eduardito, Alexander, Osmani, Alvincito; to be fair, the whole family, which also includes the two annexes (Fomento and Angel Guerra).

I appreciate a lot all the help that I received from my friends Ricardo, Ladys and Carola who patiently supported me and helped me to have a better life through these years in Japan. I'll always be in debt with you guys.

Many thanks go to my colleagues and friends: Lankesh, Lihua and all other friends from the Soshigaya International House who have helped me in so many ways. The same to my former colleagues and friends at the University of Holguin which they never gave up in their trust in me: Rosa, Félix, Rita, Mauro, Ana de Lourdes, Aleida, Velázquez, Sergio, Carmen and many others.

My doctoral studies were supported by the Japanese Ministry of Education, Culture, Sport, Science, and Technology. I thank the Japanese Government and the Japanese people in general for the unique opportunity and incredible experiences that they have given me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Thinking machines and artificial beings appear in Greek myths, such as Talos of Crete, the bronze robot of Hephaestus, and Pygmalion's Galatea [51, 77]. Human likenesses believed to have intelligence were built in every major civilization: animated cult images were worshiped in Egypt and Greece and humanoid automatons were built by Yan Shi, Hero of Alexandria and Al-Jazari [51]. By the 19th and 20th centuries, artificial beings had become a common feature in fiction. Stories of these creatures and their fates discuss many of the same hopes, fears and ethical concerns that are presented by artificial intelligence (AI). The field of AI research was founded at a conference on the campus of Dartmouth College in the summer of 1956 [51, 77].

When computers with large memories became available around 1970, researchers began to build knowledge into AI applications. This knowledge revolution led to the development and deployment of expert systems (introduced by Edward Feigenbaum), the first truly successful form of AI software [51, 77]

In the 1990s and early 21st century, AI achieved its greatest successes. At the same time, a project which aims to allow all links to any information anywhere, also had a great success, the World Wide Web (or simply the "Web"). Nowadays the main purpose of the current Web's evolution is driven to enable users to find, share and combine more easily the huge amount of information it contains. This web of data named the Semantic Web has the envision of information that can also be readily interpreted by machines. So, if machines could understand the semantics, or meaning, of information on the web,

they could perform more of the tedious work involved in finding, combining, and acting upon information on the web. The development of the Semantic Web brought some new perspectives to the Artificial Intelligence community: the "Web effect," i.e., the merge of knowledge coming from different sources, usage of URIs, the necessity to reason with incomplete data; etc.

This thesis explores an application of the semantic models to the human way of comparing words. The ability to assess similarity lies close to the core of cognition. Semantic relatedness describes the strength of the cognitive association between two concepts. For example, *man* and *woman* are very strongly related, as are *monkey* and *banana*. The concepts *screwdriver* and *truth*, however, seem to be unrelated. Other pairs of concepts often fall somewhere in between these extremes, such as *book* and *computer* or *sky-rise* and *window*. A very straightforward technique for determining the strength of relatedness between two concepts is to find the sequence of links that connects them in a semantic network [24]. The "closer" the concepts are to one another, i.e., the shorter the path that connects them, the more strongly they are related.

Early work in semantic networks proposed techniques quite similar to the shortest path length approach described above. For example, Collins and Loftus described a technique for determining semantic relatedness using the paths between nodes in a semantic network [14]. However, with the availability of WordNet - a large-scale semantic network for English [24]- a great variety of techniques for measuring semantic relatedness, and for the associated problem of measuring semantic similarity, has emerged. These new measures have provided many refinements to the approach of computing the strength of relatedness from a path in a semantic network. The goal of this study is to improve some of the existing techniques. Although our principal interest is in semantic similarity measures, many of our observations and analyses extend to semantic relatedness.

This dissertation makes several significant contributions to the study of semantic similarity. We contribute with an abstract model for words similarity assessment. Based on this model, five new semantic similarity measures were developed. In the present work the role of taxonomic properties in corpora independent metrics are also analyzed resulting

in a new metric for computing the information content of word senses. In our evaluation, we show that all the new developed measures outperformed their classical versions while one remains as competitive as their previous approach.

1.1 Aims and Motivation

The semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. The technologies of this web of data can be used in a variety of application areas; for example: data integration, knowledge representation and analysis, cataloging services, improving search algorithms and methods, social networks, etc. In order to achieve the goals of the semantic web, it have to be able to define and to describe the relations among data (i.e., resources) on the Web.

An ontology is a formal, explicit specification of a shared conceptualization. It renders shared vocabulary and taxonomy, which models a domain with the definition of entities and/or concepts, their properties and relations. They can be used to reason about the entities within that domain. Ontologies are one of the formal representation for organizing information in the semantic web and they are also used in artificial intelligence, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it. In the semantic web context, since many actors provide their own ontologies, ontology matching or ontology alignment has taken a critical role for helping heterogeneous resources to inter-operate [23].

Ontology matching tools find classes of data that are "semantically equivalent". This process determine correspondences between concepts which are called alignments [22]. Finding those correspondences imply a similarity assessment between the involved concepts. For this reason similarity measures plays an important role in ontology matching systems.

Several methods have been developed to compute words similarity, most of them operating on taxonomic dictionaries like WordNet [24] or external corpus like the Brown

Corpus [1]. However the majority of them suffer from a serious limitation. They only focus on the semantic information shared by those words, or in the semantic differences, but they have been rarely combined in a broader perspective.

1.2 Main Hypothesis and Research Questions

We will present in a synthesized manner the initial questions which the study presented in this thesis originates at.

Statement of Main Hypothesis

Because of the extensive use of words pairs comparison in several applications like ontology matching [22], information retrieval [88, 33], automatic hypertext linking [27], words sense disambiguation [75], detection and correction of malapropisms [9] and natural language processing. The central goal of our research efforts is to provide better semantic similarity measures for the task of words pairs comparison. Considering the broader object of study of words pairs comparison, our focus falls on lexical semantics similarity measures as the field of action.

Our central hypothesis can be articulated as follows.

> *The consideration of semantic commonalities and the semantic differences between words into an abstract model to be apply in the similarity computation process can improve WordNet based semantic similarity measures. Considerations of taxonomic properties should also positively impact corpus independent information content based semantic similarity measures.*

Research Questions

Several main research questions stem from the thesis stated above.

[1] The Brown University Standard Corpus of Present-Day American English (or just Brown Corpus) was compiled in the 1960s by Henry Kucera and W. Nelson Francis at Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistics. It contains 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961.

1. *Can a featured based abstract model of similarity, where semantic differences and semantic commonalities were both considered, improve semantic similarity computation process?*

2. *How will affect existing similarity measures the application of the above mentioned model?*

3. *Can corpus independent information content based similarity measures be improved if other taxonomic properties were considered?*

1.3 Contributions

The main contributions of the thesis are:

A novel model for semantic similarity computation. We show that a featured based model of similarity, where semantic differences and semantic commonalities are both considered, can be applied to word pairs comparison. We demonstrate the model application by obtaining 5 new semantic similarity measures. This work was published in conference (reviewed) [55] and in journal [57].

Five new semantic similarity measures. After applying the Menendez-Ichise model to the traditional WordNet based semantic similarity measures we obtained five new measures. We show four of this similarity measures outperformed their classical version while the last one performed the same as its' classical version. This work was published in conference (reviewed) [55] and in journal [57].

A new corpora independent information content metric. We show an analysis of taxonomic properties in corpus independent metrics. The application of this analysis allowed us to obtain a new corpora independent information content metric which generated the highest value of accuracy among the corpora dependent and the corpora independent metrics we tested. This work was published in conference (reviewed) [56] and in journal [57].

1.4 Outline and Intended Audience

The work of this thesis is organized in the following chapters.

Chapter 2- Related work discuss the related work to this thesis. We start off with an overview of the ontology matching problem and different approaches to solve it. We deepen into the similarity measures topic, describing different types of semantic similarity measures and their limitations. We finished by highlighting the reasons of our choice for applying a featured based model of similarity.

Chapter 3- A novel information content metric starts with a description of information content metrics. We describe the problematic associated to those metrics. Then we analyze the behavior of some properties in a taxonomy and their relation to the amount of content or knowledge a taxonomy's node can hold. We propose a new corpus independent information content metric. Experiments and results are also described.

Chapter 4- The Menendez-Ichise model starts describing different models of similarity from the psychology field. We applied one of them to the matching problem and we obtained the Menendez-Ichise model. Then we combine this model with traditional WordNet based semantic similarity measures obtaining five new measures. We then check the accuracy of the obtained measures. Experiments and results are also described.

Chapter 5- Conclusion and future work discusses the contribution of the thesis, gives limitations of this work and provide future directions. As future work we includes an enlargement of words' pairs dataset for estimating the best ratio between the semantic commonalities and the semantic differences, as well as the application of some machine learning methods to this ratio estimation.

The thesis above is oriented to both researchers and practitioners in the field of the semantic similarity, as well as interested readers from neighboring fields such as ontology matching, machine learning, natural language processing, etc. The theoretical results of the work are intended and have been tested on words pairs comparison. However, the findings are general and formal enough so that all the discussed approaches can be applied and/or generalized to the related fields.

Chapter 2

Related Work

The growing need of information sharing poses many challenges for semantic integration. Ontology matching, aiming to obtain semantic correspondences between two ontologies, is the key to realize ontology interoperability [22]. Recently, with the success of many on-line social networks, such as Facebook, MySpace, and Twitter, a large amount of user-defined ontologies are created and published on the Social Web, which makes it much more challenging for the ontology matching problem.

In this chapter, we provide the related work and the minimal necessary background knowledge which are important for understanding our approach to semantic similarity measures. We present the ontology matching problem (Section 2.1) and the different matching techniques and approaches described in the literature (Section 2.2). Then, we introduce the basic definitions about semantic distance (Section 2.3) and the computational resource WordNet (Section 2.4). To end the chapter we describe several relatedness (Section 2.5) and similarity measures (Section 2.6).

2.1 Ontology Matching

2.1.1 Ontology: Definitions and Applications

Defining an Ontology

Ontologies in Artificial Intelligence have been introduced to describe the semantics of data in order to provide a uniform framework of understanding between different parties. Ironically enough, despite this intention, there exists little agreement on a common definition of an ontology among different authors, many of which have proposed their own formal definitions, each taking into account different aspects of the acquisition, modeling and intended application of ontologies. The main common reference to an ontology definition was provided by Gruber back in 1993, describing ontologies as knowledge bodies which bring a formal representation of a shared conceptualization of a domain – the objects, concepts and other entities that are assumed to exist in a certain area of interest together with the relationships holding among them. Gruber wrote:

> An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence - the study of what there is. For AI systems, what exists is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge [28].

Gruber's definition suggests that an ontology possesses loosely a set of concepts and a set of relations between these concepts. The core-bodies of ontologies are taxonomies – hierarchical structures that organize concepts by a subsumption (is_a) relation. Web directories, such as *Yahoo!* or the *Open Directory Project* are examples of taxonomies which classify items in a given domain of interest.

Alexander Maedche and Steffen Staab suggested a set of characteristics that a an ontology should or could possess which they called *ontology primitives* [50]. The set of ontology primitives consists of:

1. a set of lexical entries $\mathcal{L}$ for concepts and relations;
2. a set of concepts $\mathcal{C}$;
3. a taxonomy of concepts with multiple inheritance (heterarchy) $\mathcal{H}_\mathcal{C}$;
4. a set of non-taxonomic relations $\mathcal{R}$ described by their domain and range restrictions;
5. a hierarchy (or heterarchy) on the relations $\mathcal{R}$, $\mathcal{H}_\mathcal{C}$;
6. mappings $\mathcal{F}$ and $\mathcal{G}$ that relate concepts and relations with their lexical entries, respectively;
7. a set of axioms $\mathcal{A}$ that describe additional constraints on the ontology and allow to make implicit facts explicit.

Despite the relative liberality among the members of the scientific community concerning the question *"what is an ontology?"*, Todorov provided a formal definition, which is general enough to satisfy many existing understandings of that question. Todorov did several comments on the list of primitives above, in order to clarify its components [96].

- The set $\mathcal{L}$ is understood as the set of direct lexical references to the concepts and relations in question, e.g. "School" for the concept SCHOOL, "Parent" for the relation **parent**.
- The taxonomy is defined by a partial order on the set of concepts.
- The set $\mathcal{R}$ is left without a precise definition of what kind of relations it might contain. Some examples are the partonimic relation (**part_of**), as well as non-standard relations defined by the ontology engineer (e.g. **parent**, **employed_by**, **graduated_at**, etc.).

- The set $\mathcal{A}$ includes axioms, which do not follow directly from the defined relations and concepts, but are important for modeling the respective domain. They can come from background knowledge sources like dictionaries, thesauri or top-level ontologies.

Based on the proposed list of primitives (but not considering all of its ingredients), Stumme and Maedche gave the following definition of an ontology [92]:

Definition 2.1.1. ***(Ontology)*** *A (core) ontology is a tuple* $\mathcal{O} := (\mathcal{C}, \boldsymbol{is\_a}, \mathcal{R}, \sigma)$*, where* $\mathcal{C}$ *is a set whose elements are called concepts,* $\boldsymbol{is\_a}$ *is a partial order on* $\mathcal{C}$*,* $\mathcal{R}$ *is a set whose elements called relation names (or relations for short), and* $\sigma : \mathcal{R} \rightarrow \mathbb{N}$ *is a function which assigns to each relation name its arity.*

In the next sub-section, we will consider various ontology application scenarios.

Ontology Applications

"Ontology applications" is the title of the last and most voluminous part of the *Handbook of Ontologies*, edited by Stefan Staab and Rudi Studer [89]. Its eleven chapters go into different aspects of application of ontologies in multiple real life scenarios. Since this is one of the few available endeavors in classifying ontology application fields to date, we will describe it briefly in the sequel. Todorov argued that the application fields can be organized in a more consistent manner [96].

Staab and Studer have conventionally classified the ontology application fields into two big families – *Knowledge Management* and *Interoperability and Integration (of Enterprise Applications)* (Figure 2.1).

The first class of applications aims at answering the question how ontologies can be of help in support of the identification, creation, representation and distribution of knowledge. This includes the use of ontologies to support the corporate memory of a virtual business partnership using flexible, but well understood for both humans and machines document-based data structures. Ontologies are main bodies in the Semantic Web, therefore researchers and practitioners have put efforts into developing different

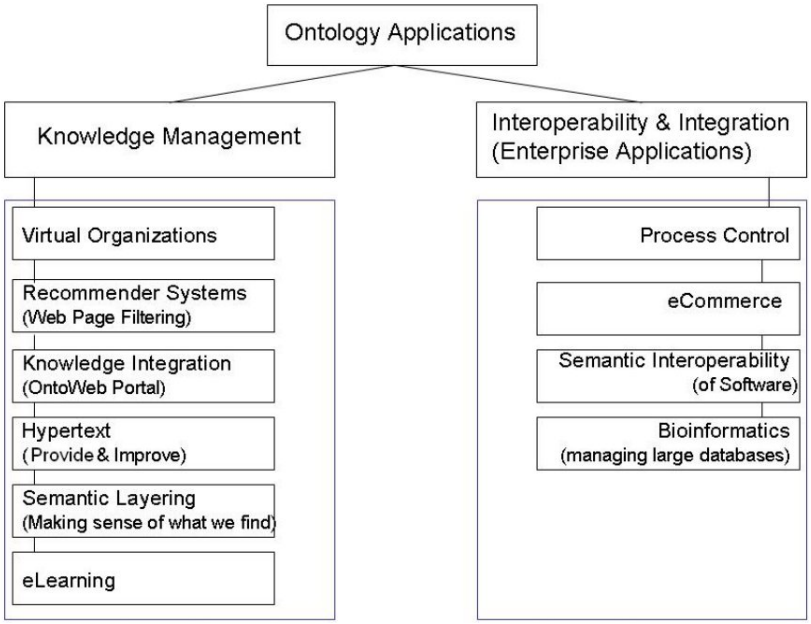

Figure 2.1: *Ontology applications: the classification of Staab and Studer.*

Semantic Web improvement scenarios [7]. In that context, ontologies are applied in various topics, such as *Recommender Systems* (Web Page Filtering), *Knowledge Integration* (the OntoWeb Portal project[1]), provision and improvement of *hypertext*, *Semantic Layering* (or "making sense of what we find"). Finally, using ontologies to support *eLearning* finds ultimately place in this general class of applications.

The second class of applications is centered around the role of ontologies for providing interoperability and integration of enterprise applications. It discusses applications in the fields of *Process Control* (within a company or between multiple partner companies), *semantic interoperability of software* (how to enable the cooperation of two software applications that were initially not developed for this purpose) and *eCommerce*. Finally, ontologies are able to manage large data bases; this has found place in Staab and Studer's classification in the context of *bio-informatics* – a broad contemporary ontology application field.

However, Todorov make some comments on the structure and completeness of the previous classification [96].

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

- Provided the big variety of real life areas where ontologies play a role, splitting the various application fields in only two classes is not granular enough.

- Moreover, the reader is left with the impression that most applications of both classes at the end have to do with knowledge management for the purposes of the *eBusiness*, which is a false suggestion.

- Virtual Organization, Knowledge Integration and Semantic interoperability in their essence tackle with the same problem and are driven by the same motivation of providing a mutual framework for semantic homogeneity and aim at similar application domains.

- Semantic Layering, Hypertext and Recommender Systems can also be grouped together because, as observed above, they are basically Semantic Web driven applications.

- Interoperability and Integration of data may be viewed as a sub-domain of Knowledge management.

- Some important application fields have been left out.

  - Ontologies in support of *problem solving* is a prominent application domain. Problem solving methods provide reusable reasoning components by specifying the way in which new facts can be inferred from existing facts on the basis of some set of logical axioms complementing an ontology. In that sense they could be classified in the Knowledge Management part of ontology application tree or they can form a class of their own – ontologies as inference systems.

  - Planning in Artificial Intelligence deals with building action strategies to be realized by intelligent agents. Applying ontologies in planning for providing semantics to the sequences of actions is a growing research topic.

  - Ontologies play an important role in *Natural Language Processing*. Estival et al. discuss the possibility of coupling ontologies with the lexicon used in a natural

language component of a system for facilitating presenting and retrieving of information [21].

Todorov's contribution to a classification of the ontology application domains builds on the work of Mizogouchi and Staab and Studer by using parts of the typology presented by Mizogouchi in order to classify the application domains discussed by Staab and Studer and some more fields that we find necessary to include [60, 89]. The new classification tree is represented in Figure 2.2.

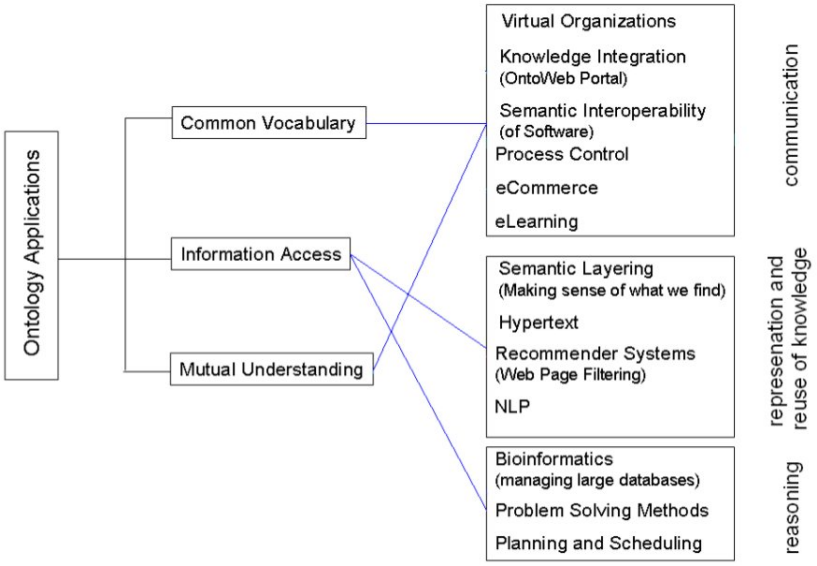

Figure 2.2: *Application fields of ontologies: the classification of Todorov.*

Todorov splits the ontologies applications in three main blocks, containing intersecting application instances [96].

- **Ontologies providing a common vocabulary**

  This is the most intuitive and straightforward type of ontology application: having a common vocabulary is the first step towards the systematization and the sharing of knowledge of a given domain.

- **Ontologies in support of information access**

  Ontologies provide vocabulary for annotation of web resources and enable agents to use hierarchy and class relations in order to interpret this vocabulary. This is

a step towards making information access more intelligent and using the enormous information sources of the World Wide Web.

- **Ontologies for mutual understanding**

  Mutual understanding considers two types of communicating agents humans and software agents. Each of them can be on either side of the communication line.

  Communication between humans can be facilitated by ontologies by providing environments for knowledge-intensive engineering such as concurrent engineering, business process re-engineering and other.

  A big part of the ongoing ontology research driven by the core ideas that lie in the project of the Semantic Web concerns *understanding between humans and software agents*, seen in the case of web resources search and use [7].

  Communication between software agents has been discussed above in terms of allowing the cooperation of two software applications that were initially not developed for this purpose.

2.1.2 Motivations of Ontology Matching

In the broadest sense, ontology matching is the process of finding correspondences between the elements of two or more heterogeneous ontologies. The following sections introduce phenomenas related to the possible ambiguities emerging among communities in representing semantic knowledge, which underlie the problem of ontology heterogeneity [96]. Since semantic similarity is introduced to re-establish the links between different conceptual representations of the same entities and thus provides basic building blocks for an ontology matching procedure, similarity, which is the main topic of this research, will be discussed later on in relation to human concept formation and measuring semantic proximity of concepts.

Representation of Semantic Knowledge

In order to explain the problem of ontology matching, its motivation and possible solutions, we need to examine more deeply the representation of semantic knowledge and a phenomenon called semantic heterogeneity. Ogden and Richards, back in 1923, described the relation between the real world objects, the concepts (defined most commonly as mental representations) and symbols (language expressions) introducing the so called meaning triangle: a symbol stands for a real world object and evokes a concept; a concept refers to a real world object. Later on, in 2000 John Sowa built on top of the meaning triangle the knowledge representation triangle, aiming to show how a person connects his concept with a certain conceptual representation [87]. On the knowledge representation triangle's vertices we find: the concept (a vertex from the meaning triangle), a representation of a concept (which models the concept) and a concept of representation (which relates to the concept of the representation of a concept). Figure 2.3 summarizes these ideas.

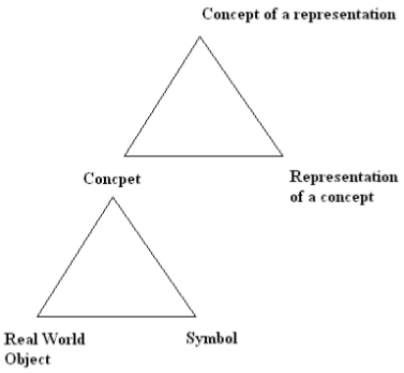

Figure 2.3: *The meaning triangles of Ogden, Richards and Sowa.*

Clearly, the mental representation (concept) and the choice of a symbol (word) for a given real world object may differ among different people. For example, the symbol for a given real world object, say an electric guitar amplifier, may differ because of the language different people choose to use. "Amplifier" might be the symbol chosen by an English speaking person, while it is more likely that a Spanish native speaker chooses the symbol "Amplificador" for the a same real world object. More over, even among people that have reached an agreement on the use of a common natural language, the symbolic ambiguity

might still appear. It is possible that a guitar player that works everyday with electric guitar amplifiers and knows a lot about different kinds of amplifiers (for which reason he needs to distinguish between them) might call it a "Squire" (following the brand of his favorite amplifier manufacturer), someone else might still use the symbol "Amplifier" or the abbreviation "Amp."

The conceptual ambiguity appears quite often, too, since different people develop different mental representations of one and the same set of real world objects, depending on the categorization principles they decide to use and the references to the category they have among the real world objects. It is argued that this is a complex process in the core of which lie various historical and cultural conditions, as well as complex psychological processes. Particularly importance has the question about the role similarity plays in this whole process and how is it defined and perceived; especially because similarity of concepts helps us judge on their semantic proximity.

Finally, going up to the knowledge representation triangle, we will find different representations of one and the same real world object, since its conceptualization varies among different people. Todorov claimed that this is where ontology heterogeneity evolves from [96]. Solving the heterogeneity problem by the definition and application of various measures of (semantic) similarity of concepts introduces rules for translation between different conceptual systems – essential for the mutual understanding and interoperability between (human or artificial) agents.

Ontology Heterogeneity: Aspects of the Problem of Ontology Matching

Ontologies, as knowledge representation bodies, suffer the problem of knowledge representational heterogeneity described above for many reasons, mostly because of the limitations following from the decentralized and strongly human-biased nature of ontology acquisition. Ontologies are being created from different people and communities independently from one another and this process is largely manual or, in the best case, semi-automatic. In many open and evolving systems and applications with decentralized

nature where ontologies are broadly applied, such as Peer-2-Peer Systems, eCommerce or the widely discussed Semantic Web, it is unlikely that different parties would adopt the same ontologies to represent the same fragments of knowledge [7]. This has lead to the creation of a considerable number of ontologies, which describe similar or overlapping domains of knowledge but their elements do not explicitly match – a phenomenon called *ontology heterogeneity* [96]. (See Figure 2.4 for an illustration of two ontologies which share a semantic overlap).

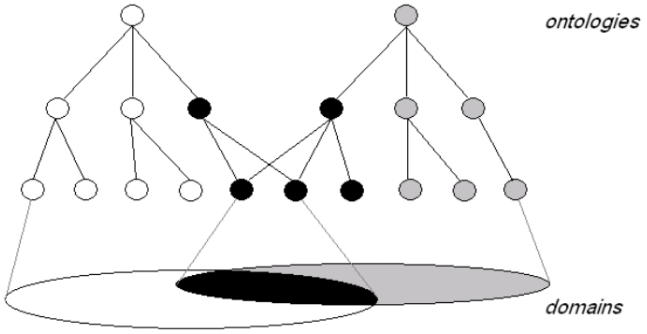

Figure 2.4: *Overlapping ontologies.*

Ontology matching amounts to reducing ontology heterogeneity and, ultimately, overcoming the barriers in front of knowledge sharing. In the beginning of the section, we spoke of the problem of heterogeneity in knowledge representation. Ontology heterogeneity originates at that problem and can occur in many different forms. Many authors have provided classifications of the different types of mismatches that can occur among schema, databases and ontologies [5, 22, 42, 85]. Following the heterogeneity typology by Euzenat and Shvaiko, one distinguishes between four main types of heterogeneity and we will briefly describe each of them [22].

Syntactical heterogeneity concerns ontologies, which are expressed in different formal languages [3]. As argued by many authors, this type of heterogeneity is among the easiest to overcome. It should be tackled on a theoretical level by defining correspondences between the constructs of the different languages.

Terminological heterogeneity is about vocabulary mismatch: differences in the choices

of names when referring to the same ontological entities (concepts, relations or instances). Usually, lexical or instance-based matching techniques are applied in order to find correspondences between such entities.

Conceptual heterogeneity refers to three sub-types of differences when modeling the same domain of interest:

- Differences in coverage: two ontologies describe different or partly overlapping domains, from the same perspective and in the same detail;
- Differences in granularity: two ontologies describe the same domain, from the same perspective, but in different details;
- Differences in scope: two ontologies describe the same domain with the same level of detail, but from a different perspective.

Finally, *semiotic* or *pragmatic heterogeneity* collects mismatches in how entities are interpreted by people in a given context and is hard to model computationally. Of course, the typology presented above is not universal and the different heterogeneity types often appear simultaneously. For instance, syntactical heterogeneity, as described here, can result in semantic differences; terminological differences, on the other hand, are also considered in various sources as syntactical.

As already observed, reducing heterogeneity is achieved in terms of identifying similarity. More generally speaking, the match as an operation on structured information can be defined as an operation, which takes two ontologies as an input and produces a mapping between those elements of the two ontologies that correspond semantically to each other. Therefore, the task of ontology matching can be viewed as the identification of similarities between the different elements of two distinct ontologies, by applying a defined distance function or measure of similarity between two ontologies or their alignable elements[2] [31]. A very general definition of this process can look like that:

> Ontology matching is the process of identifying the implicitly contained similarities between the elements of two heterogeneous ontologies, which cover

[2]By alignable elements Todorov means elements that correspond to the same component of an ontology definition.

the same or similar domains of knowledge but their elements do not explicitly match.

A schematic representation of the ontology matching process is given in Figure 2.5. The ontology matching processes have two main stages: the matching stage and the user interaction stage. Several matching algorithms/techniques can be used during the matching stage of the matching process [9, 22, 55]. The second stage gives a taste of the iterative property which characterized all ontology matching processes and it creates an ideal scenario for the application of relevance feedback techniques.

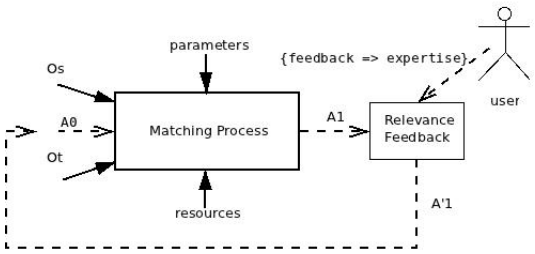

Figure 2.5: *General framework for the ontology matching process.*

2.2 Ontology Matching Techniques and Approaches

There are two classifications of ontology matching approaches. The first one is provided by Rahm and Bernstein (2001) and is among the most general available frameworks [73]. The second one builds on the first one and was published by Euzenat and Shvaiko in their ontology matching book (2007) [22].

Euzenat and Shvaiko [22], provided a very detailed classification of ontology matching approaches, which builds on Rahm and Bernsteins taxonomy, but is more exhaustive and granular, taking into consideration the advances made in this dynamic field in the time gap between the two publications. In building their classification, the authors have used as guidelines four main criteria: exhaustivity (the sub-categories of a given category all together contain exactly the extension of that category), pair-wise disjointness of the categories, homogeneity of classes and, finally, adding and modifying classes until saturation has been reached.

In Figure 2.6, we have presented a fragment of Euzenat and Shvaikos classification which we consider to be granular enough for the purposes of our study and for giving a ground for situating our approach and related methods. We will discuss in more details each of the different categories of methods by putting a closer focus on approaches which are directly relevant to ours.

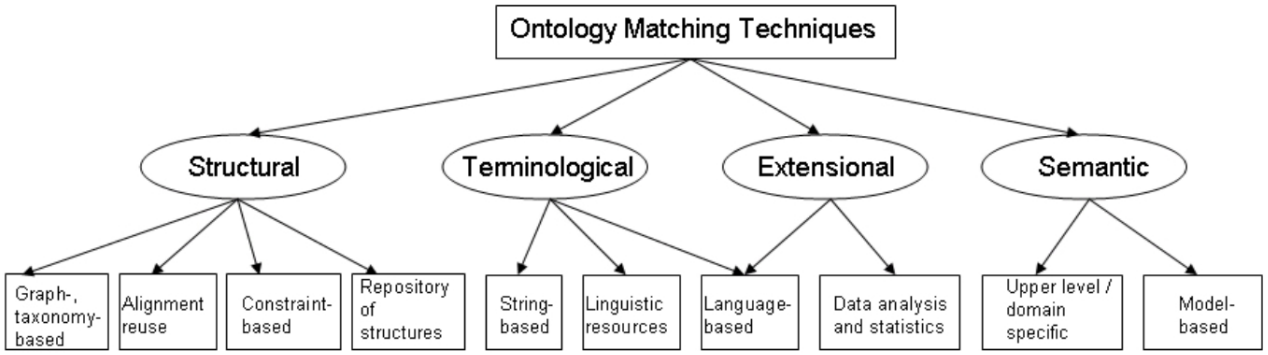

Figure 2.6: *Matching approaches (a fragment from [22]).*

2.2.1 Structural Approaches

The structure of an ontology can be studied on two different levels with respect to either a single ontology element, known as internal structure, or the way in which a set of elements are related, known as relational structure. Methods based on the former structure type look into similarities of the sets of properties of two elements, the data types used to describe them or their properties, the cardinalities that sets of values of two properties are "allowed" to reach, etc. These approaches are suited to schema matching problems where one disposes readily with an internal structure of the database entities.

Other approaches, do not rely on the explicit availability of such structural information, but they are concerned with comparing blocks of elements together with the relations that hold between them. This is exactly the focus of relational structure based methods; the matching problem is typically situated in a graph-theoretical framework where an ontology is modeled as a graph whose vertices and edges are labeled by concepts and relations names, correspondingly. The matching of such "ontology graphs" is usually re-

duced to solving a graph isomorphism problem and identification of a maximal common sub-graph of two graphs.

Following the classification of graph matching problems given by Bengoetxea [6], we distinguish between *exact* and *inexact* matching. Exact matching is defined as the graph matching problem when there exists an isomorphism[3] from one graph to another or from a sub-graph of a graph to (a subgraph of) another graph. The term inexact matching is used to denote a class of matching problems for which it is not possible to find an isomorphism between the two input graphs. It consists in finding the *best* possible matching between the vertices of two graphs, rather than the *exact* node-to-node correspondence.

The algorithms available for solving graph matching problems can be classified into *optimal* and *suboptimal*. The problem of finding a maximal common subgraph, which underlies most graph matching algorithms, is NP-complete. A couple of state-of-the-art graph matching algorithms and approaches designed for various application fields provide an *optimal* solution in exponential time and space which makes them computationally intractable. On the other hand, *suboptimal* or *approximative* methods are able to find a solution in polynomial time, but give no guarantee that the solution found is not due to a local minimum trap [1]. For a more advanced discussion of the overall problem of graph matching in terms of theoretical foundations, algorithms and applications, we refer to [6, 11].

Some systems which use an structural approach are: Anchor-Prompt [63], Cupid [49], Onion tool [59] and Chimaera tool [53].

2.2.2 Extensional Approaches

Extensional ontology matching, also known as instance-based matching, comprises a set of theoretical approaches and tools for aligning two or more heterogeneous ontologies based on their extensions – the instances that populate their concepts. In hierarchically structured data, a concept can be defined as a set of those instances that are directly

[3]In graph theory, an isomorphism of graphs $\mathcal{G}$ and $\mathcal{H}$ is a bijection between the vertex sets of $\mathcal{G}$ and $\mathcal{H}$, $f : \mathcal{V}(\mathcal{G}) \rightarrow \mathcal{V}(\mathcal{H})$ such that any two vertices u and v of $\mathcal{G}$ are adjacent in $\mathcal{G}$ if and only if $f(u)$ and $f(v)$ are adjacent in $\mathcal{H}$.

assigned to it, or as a set of all instances assigned to the concept and its successors in the taxonomy, what we will later call *hierarchical* and *non-hierarchical* instantiation. An example of a hierarchical instantiation is given in Figure 2.7, D1, D2 and D3 are the sets of instances of the leaves of the tree. An "instance" can be considered as some kind of a real-world data entity, a member of a given class within an ontology.

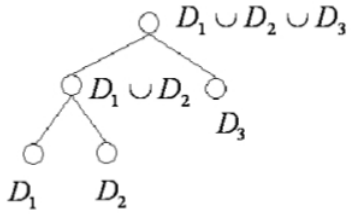

Figure 2.7: *Hierarchical instantiation of a taxonomy (adapted from [96]).*

A common approach to modeling concepts by their instances is the set-theoretic approach. The relatedness of a pair of concepts is an outcome of a properly chosen measure of similarity, based on estimations of the intersections of two sets of instances. Some systems which use this approach can be found in [19, 44, 92, 93, 95, 100, 101].

2.2.3 Semantic-based Approaches

The group of semantic-based approaches unites methods, which rely on logical deduction in order to justify and verify a set of previously generated mappings. As this definition suggests, these methods usually consist of two main parts:

1. **Anchoring the source ontologies**. This is the process of initial alignment of two (or more) source ontologies by matching them against an existing external resource of some kind. This can be a formal top-level ontology, such as DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) [25] and SUMO (Suggested Upper Merged Ontology) [66], a formal domain specific ontology like, for instance, the FMA (Formal Model of Anatomy) ontology in the medical domain, or an informal resource, like WordNet [24] . The main characteristic of this method is that the input ontologies (the mapping candidates) are first aligned to (parts

of) the background ontology. Checking if the concepts and relations of the source ontologies correspond to one another is performed by the help of reasoning services in the background ontology.

2. **Applying deductive techniques**. The second part of the mapping procedure consists of verifying the consistency and the completeness of the correspondences found in the first phase and entailment of new alignments. To these ends, on applies techniques from propositional or description logics for verifying semantic satisfiability of the correspondences and deduce new knowledge.

Works related to semantic-based approaches can be found in [22, 39, 34, 64, 65].

2.2.4 Terminological Approaches

Terminological methods comprise two major groups of approaches – those that use strings in order to match names of entities (lexical similarity measures or string based similarity measures), and those that rely on linguistic information contained in dictionaries and thesauri combined with techniques from Natural Language Processing in order to compare the similarity of terms and their relations and overcome problems evolving from *synonymy* and *polysemy* (lexical semantic similarity measures or simply semantic similarity measures).

In the next subsection we will introduce the most used lexical similarity measures. Since the semantic similarity measures are the main target of this study we will analyze them separately.

Intuitively, similarity between two objects can be measured in terms of their distance in a certain (metric) space or in terms of shared common features. We will start by defining and distinguishing between a distance function and a similarity (dissimilarity) measure, which are core notions for the similarity problem.

Definition 2.2.1. ***(Metric or Distance)*** *A* ***metric*** *on a set $\mathcal{X}$ is a function (called the distance function or simply* ***distance****) $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, (where $\mathbb{R}$ is the set of real*

numbers). For all x, y, z in $\mathcal{X}$, this function is required to satisfy the following conditions:

$$
\begin{aligned}
&d(x,y) \geq 0 && (\textit{non-negativity})\\
&d(x,y) = 0 && \textit{if and only if } x = y \textit{ (identity of indiscernibles)}\\
&d(x,y) = d(y,x) && (\textit{symmetry})\\
&d(x,z) \leq d(x,y) + d(y,z) && (\textit{subadditivity / triangle inequality})
\end{aligned}
$$

In computer science, the notion of distance is sometimes considered as weaker and thus not synonymous to a metric, in contrast to the pure mathematical sense of this notion. Example is the edit distance, to be discussed below, which is not a metric in the general case. However, throughout the thesis by metric and distance function we will refer to the same concept.

The notion of similarity or dissimilarity is more liberal than that of a metric. Dissimilarity is intuitively related to the distance between two entities, whereas similarity is the exact inverse of dissimilarity. To assess these concepts usually a measure of dissimilarity is defined which relaxes one or more of the conditions for a distance function. We will introduce the most commonly adopted definitions of similarity and dissimilarity between two entities [18].

Definition 2.2.2. ***(Dissimilarity)*** *Let $\mathcal{X}$ be a set and let $x, y \in \mathcal{X}$. A **dissimilarity** function on the set $\mathcal{X}$ is defined as a mapping $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:*

$$
\begin{aligned}
&\delta(x,x) = 0 && (\textit{minimality})\\
&\delta(x,y) \geq 0 && (\textit{non-negativity})\\
&\delta(x,y) = \delta(y,x) && (\textit{symmetry})
\end{aligned}
$$

Definition 2.2.3. ***(Similarity)*** *Let $\mathcal{X}$ be a set and let $x, y, z \in \mathcal{X}$. A **similarity** function on the set $\mathcal{X}$ is defined as a mapping $\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:*

$$
\begin{aligned}
&\sigma(x,x) \geq \sigma(y,z) && (\textit{minimality})\\
&\sigma(x,y) \geq 0 && (\textit{non-negativity})\\
&\sigma(x,y) = \sigma(y,x) && (\textit{symmetry})
\end{aligned}
$$

Lexical Similarity Measures

There are several lexical similarity measures (also known as string-based methods) in the ontology alignment field although as the reader has seen they are widely used in several fields and applications. These techniques focus on entity's name (string) and find similar string entities. Here, we briefly introduced the most popular methods which are already implemented in Alignment API [16] and SecondString API [12, 13].

- **N-gram similarity** compares two strings and calculates the number of common n-grams between them. An n-gram is composed of all sequences of n characters [43]. For instance, three-gram of word "paper" are: "pap", "ape" and "per".

- **Levenshtein distance** is a metric for measuring the amount of difference between two sequences (i.e. an edit distance). The term edit distance is often used to refer specifically to Levenshtein distance. The Levenshtein distance between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character. For example, the Levenshtein distance between "kitten" and "sitting" is 3, since the following three edits change one into the other, and there is no way to do it with fewer than three edits:

  1. ***k***itten ***s***itten (substitution of 's' for 'k')
  2. sitt***e***n sitt***i***n (substitution of 'i' for 'e')
  3. sittin sittin***g*** (insertion of 'g' at the end).

  In the levenshtein distance different penalty costs can be given to insertion, deletion and substitution operations. We can also give penalty costs that depend on which characters are inserted, deleted or substituted.

- **SMOA distance** is based on the number of common parts in two strings, while considering the length of mismatched substrings and the length of the common prefix in both strings [90].

- **Dice coefficient** is defined as twice the number of common terms of the compared strings over the total number of terms in both strings. The coefficient result of 1 indicates identical vectors, while 0 equals orthogonal vectors. Dice coefficient, is a similarity measure related to the Jaccard index [13].

  For sets X and Y of keywords used in information retrieval, the coefficient may be defined as twice the shared information (intersection) over the sum of cardinalities:

$$s = \frac{2 * |X \cap Y|}{|X| + |Y|} \tag{2.1}$$

  When taken as a string similarity measure, the coefficient may be calculated for two strings, x and y using bi-grams as follows:

$$s = \frac{2n_t}{n_x + n_y} \tag{2.2}$$

  where n_t is the number of character bi-grams found in both strings, n_x is the number of bi-grams in string x and n_y is the number of bi-grams in string y.

  For example, to calculate the similarity between: *night* and *nacht*. We would find the set of bi-grams in each word: $\{ni, ig, gh, ht\}$ and $\{na, ac, ch, ht\}$ Each set has four elements, and the intersection of these two sets has only one element: *ht*. Inserting these numbers into the formula, we calculate, $s = (2 * 1)/(4 + 4) = 0.25$.

- The **Jaccard index**, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A, B) = \frac{A \cap B|}{|A \cup B|} \tag{2.3}$$

  The **Jaccard distance**, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard

coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union [90].

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{2.4}$$

- **Jensen-Shannon distance** is a popular method of measuring the similarity between two (or more) probability distributions [13, 54].

  Let X be a discrete random variable that takes on values from the set $\mathcal{X}$ with probability distribution $p(x)$. The (Shannon) entropy of X is defined as:

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2.5}$$

  The relative entropy or Kullback-Leibler(KL) divergence between two distributions $p_1(x)$ and $p_2(x)$ is defined as:

$$KL(p_1, p_2) = \sum_{x \in \mathcal{X}} p_1(x) log \frac{p_1(x)}{p_2(x)} \tag{2.6}$$

  KL-divergence is a measure of the "distance" between two probability distributions; however it is not a true metric since it is not symmetric and does not obey the triangle inequality. KL-divergence is always non-negative but can be unbounded; in particular when $p_1(x) \neq 0$ and $p_2(x) = 0$, $KL(p_1, p_2) = \infty$. In contrast, the Jensen-Shannon(JS) divergence between p_1 and p_2 defined by:

$$\begin{aligned} JS_\pi(p_1, p_2) &= \pi_1 KL(p_1, \pi_1 p_1 + \pi_2 p_2) + \pi_2 KL(p_2, \pi_1 p_1 + \pi_2 p_2) \\ &= H(\pi_1 p_1 + \pi_2 p_2) - \pi_1 H(p_1) - \pi_2 H(p_2), \end{aligned} \tag{2.7}$$

  where $\pi_1 + \pi_2 = 1$, $\pi_i \geq 0$, is a measure that is symmetric in $\{\pi_1, p_1\}$ and $\{\pi_2, p_2\}$, and is bounded.

- **Monge-Elkan distance** uses semantic similarity of a number of strings or substrings. Each substring is evaluated against the most similar substring in the com-

parison entity names [91, 71].

Let us assume that the strings s and y are broken into substrings (tokens), i.e., $s = s_1 \dots s_K$ and $y = y_1 \dots y_L$. The intuition behind Monge-Elkan measure is the assumption that s_i in s corresponds to a y_j with which it has highest similarity. The similarity between s and y equals the mean of these maximum scores. Formally, the Monge-Elkan (ME) metric is defined as follows, where sim denotes some secondary similarity function.

$$ME(s, y) = \frac{1}{K} \sum_{i=1}^{K} \max_{j=1\dots L} sim(s_1, y_j) \tag{2.8}$$

- **Substring similarity** calculates the similarity of two strings based on their common longest substring [22].

  Substring similarity is a similarity $\sigma : \mathcal{S} \times \mathcal{S} \to [0,1]$ such that $\forall x, y \in \mathcal{S}$, and t be the longest common substring of x and y:

$$\sigma(x, y) = \frac{2|t|}{|x| + |y|} \tag{2.9}$$

- **Needleman-Wunsch** applies a global alignment on two sequences (strings). It is a suitable measure when the two sequences are of similar length, with significant degree of similarity throughout. It also determines whether it is likely that two sequences evolves from the same string [12, 16].

- **Smith-Waterman distance** is a version of Needleman-Wunsch which measures the local sequence alignment. In other words, it determines similar regions between two string sequences. Instead of looking at the total sequence, this algorithm compares segments of all possible lengths and optimizes the similarity measure [13].

- **Cosine similarity** transforms the input string into vector space so that the Euclidean cosine rule is used to determine similarity [90].

  Given two vectors of attributes, A and B, the cosine similarity, θ, is represented as:

$$similarity = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{2.10}$$

The resulting similarity ranges from 1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

For text matching, the attribute vectors A and B are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

- **Jaro distance** (d_j) finds words with spelling mistakes.

$$d_j = \frac{1}{3} \cdot \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \tag{2.11}$$

where:

  - m is the number of matching characters;
  - t is half the number of transpositions;
  - and two characters from string s_1 and string s_2 respectively, are considered *matching* only if they are not farther than $\lfloor \frac{\max(|s_1|,|s_2|)}{2} \rfloor - 1$.

The **JaroWinkler distance** is a measure of similarity between two strings. It is a variant of the Jaro distance metric and mainly used in the area of record linkage (duplicate detection). The JaroWinkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match [17].

The JaroWinkler distance uses a prefix scale p which gives more favorable ratings to strings that match from the beginning for a set prefix length l. Given two strings s_1 and s_2, their JaroWinkler distance (d_w) is:

$$d_w = d_j + (l_p(1 - d_j)) \tag{2.12}$$

where:

- d_j is the Jaro distance for strings s_1 and s_2;
- l_p is the length of the common prefix at the start of the string up to a maximum of 4 characters;
- p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. p should not exceed 0.25, otherwise the distance can become larger than 1. The standard value for this constant in Winkler's work is $p = 0.1$.

2.3 Semantic Distance

The notion of *semantic distance* – sometimes called *conceptual distance* – has received a great deal of attention in the field of lexical semantics in recent years. In general, semantic distance denotes the degree of semantic association between concepts. However, many authors, including Resnik, Budanitsky and Hirst distinguish two kinds of semantic distance: *semantic similarity* and *semantic relatedness* [74, 10]. Whereas similarity expresses the degree to which two concepts resemble one another, relatedness encompasses a wide variety of semantic relationships.

Although semantic similarity and semantic relatedness have received the most study, these senses do not exhaust the range of possible types of semantic distance. For example, Budanitsky and Hirst argue that distributional similarity describes a phenomenon that is distinct from both semantic similarity and semantic relatedness [10]. However, Scriver introduced another sense of semantic distance called semantic contrast which differs from both similarity and relatedness in important ways [82].

The current study is concerned primarily with semantic similarity, which has been argued that in many cases is an adequate proxy for relatedness. In fact, in a study by Budanitsky and Hirst that evaluated the performance of a number of similarity and relatedness measures for relatedness tasks, the authors found that similarity measures achieved better results than the relatedness measures [10]. However, in this chapter,

we will therefore review both relatedness and similarity measures, including all of the measures compared by Budanitsky and Hirst. Two other promising measures that were not included in Budanitsky and Hirst's study will also be described, including one by Yang and Powers [103] and another by Banerjee and Pedersen [2].

Semantic distance measures have been developed using a variety of lexical resources. However, the scope of this study will be mainly limited to measures that employ the WordNet lexical database. There are two reasons for restricting the study to only WordNet-based measures. First, as all of the measures to be compared share a common primary resource, the validity of comparisons between the measures will not be compromised by the quality of the lexical resources that they use.

Second, most of the major approaches to measuring either similarity or relatedness are represented by WordNet-based measures. The notable exception to this are measures that employ corpus statistics to determine distributional similarity. Such measures rely on the observation that words that occur in similar contexts are likely to be semantically similar. Mohammad and Hirst provide a theoretical comparison between corpus-based measures of distributional similarity and taxonomy-based relatedness and similarity measures, and conclude that an experimental comparison is also required [61].

However Mohammad and Hirst also conclude that to a certain extent the two types of measure are incommensurable. While taxonomy-based approaches measure the similarity of concepts, corpus-based approaches measure the similarity of words. Mohammad and Hirst suggest that it may be more reasonable to view distributional similarity as a phenomenon distinct from conceptual similarity. As a result of these concerns, corpus-based measures of distributional similarity are excluded from the scope of this study.

The scope of this study is also limited to measures of the semantic distance between lexicalized concepts, which is to say, concepts that are expressed by individual words in the English language. Insofar as the primary use of semantic distance measures lies in natural language processing, lexicalized concepts deserve the most attention from a practical point of view. For the rest of this study, any reference to "concepts" may be assumed to refer specifically to "lexicalized concepts". To avoid redundancy, the terms

"lexical" and "semantic" will often be dropped so that, for example, "lexical semantic similarity" will be simply "similarity."

2.3.1 Similarity and Relatedness

Although the difference between lexical semantic similarity and lexical semantic relatedness can sometimes be subtle, it is nevertheless significant. Similarity can be understood to denote a kind of familiar resemblance. It is sometimes described in terms of featural overlap [99]. Under this view, the similarity of two concepts is the degree to which they share features in common. Features that are common to two concepts indicate greater similarity, and features that are peculiar to one or the other indicate reduced similarity. In this study we are not committed to a feature-based representation of concepts, but features provide a useful way of talking about similarity.

In contrast to similarity, relatedness describes the degree to which concepts are associated via any kind of semantic relationship. These relationships can include the classical lexical relations such as synonymy, hypernymy (*IS-A*), and meronymy (*HAS-A*), and also what Morris and Hirst have called "non-classical relations [62]." In fact, even the relation of similarity is encompassed by relatedness. As a result, all similar concepts are also related – by virtue of their similarity – such that similarity may be viewed as a special case of relatedness.

The difference between similarity and relatedness is often illustrated with examples. Resnik provides the widely used example of *car* and *gasoline* [74]. Cars and gasoline are not very similar; they have very few features in common. Whereas a *car* is a solid mechanical device, *gasoline* is a combustible liquid. An itemization of the properties of cars and gasoline would have little overlap. In spite of their differences, however, *car* and *gasoline* are very closely related through their functional association, namely that cars use gasoline. Thus, while in terms of similarity *car* and *gasoline* are semantically distant, in terms of relatedness they are semantically close.

2.4 Computational Resources

2.4.1 WordNet

All of the computational measures of semantic distance that will be discussed in this study employ the WordNet lexical database [24]. WordNet is a lexical reference system that was created by a team of linguists and psycholinguists at Princeton University. The purpose of WordNet is to model the English lexicon according to psycholinguistic theories of human lexical memory. WordNet may be distinguished from traditional lexicons in that lexical information is organized according to word meanings, and not according to word forms. As a result of the shift of emphasis toward word meanings, the core unit in WordNet is something called a synset. Synsets are sets of words that have the same meaning, that is, synonyms. A synset represents one concept, to which different word forms refer. For example, the set car, auto, automobile, machine, motorcar is a synset in WordNet and forms one basic unit of the WordNet lexicon. Although there are subtle differences in the meanings of synonyms – often differences of connotation rather than of denotation – these are ignored in WordNet. Table 2.1 and Table 2.2 show some statistics and polysemy information on WordNet 3.0.

Table 2.1: *WordNet 3.0 statistics: number of words, synsets, and senses.*

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Table 2.2: *WordNet 3.0 statistics: polysemy information.*

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101863	15935	44449
Verb	6277	5252	18770
Adjective	16503	4976	14399
Adverb	3748	733	1832
Totals	128391	26896	79450

WordNet synsets are linked together by semantic relations to form a network. These relations include hypernymy (*IS-A*) and meronymy (*HAS-A*), among others. Some relations that hold between word forms have also been included in WordNet, such as derivational relatedness. WordNet synsets are divided into nouns, adjectives, verbs, and adverbs. Although there is some interconnectivity between the different speech categories, it is quite limited. The portions of WordNet for each part of speech also have different properties, and may therefore require special treatment. For example, while the hypernymy relation is central to the organization of the noun portion of WordNet, adjectives are organized primarily in terms of the antonymy and similarity relations. Although the current WordNet version is 3.0, Table 2.3 provides a complete list of WordNet 2.0 relations and their frequency count by category [83].

Table 2.3: *WordNet 2.0 relations and frequency count by type, reproduced from [83].*

Relations	Noun	Adjective	Verb	Adverb
Antonym	2074	4118	1079	722
Hypernymy (*IS-A*)	81857		12985	
Hyponym (*SUBSUMES*)	81857		12985	
Member holonym (*PART-OF*)	12205			
Substance holonym	787			
Part holonym	8636			
Member meronym (*HAS-A*)	12205			
Substance meronym	787			
Part meronym	8636			
Attribute	648			
Derivation	21491		21497	3209
Category domain	3789	1125	1215	37
Category member	6166			
Region domain	1200	76	2	2
Region member	1280			
Usage domain	654	237	18	74
Usage member	983			
Entailment			409	
Cause			218	
Verb group			1748	
Similar to		22196		
Participle of verb		124		
Pertainym		4711		
Also see		2697	597	
Totals	245255	35932	52753	4044

Many of the similarity measures discussed in this study apply to nouns exclusively

and rely closely on the special properties of the noun subgraph of WordNet. The primary organizing relations in the noun part of WordNet are hypernymy and hyponymy. A concept is a hyponym if it is a specific type of a more general class. For example, a *robin* is a kind of *bird* and is therefore a hyponym of *bird*. The inverse of a hyponym is a hypernym, which denotes a more general class with respect to a more specific one. Thus *bird* is a hypernym of *robin*. Part/whole relations, including meronymy and holonym, also play an important role in the noun portion of WordNet. A concept is a meronymy if it is part of a whole, whereas a concept is a holonym with respect to its constituent parts. However, nearly 80% of semantic relations between nouns are hypernymy or hyponymy [10]. The hierarchical nature of the is-a relation results naturally in a tree-like structure. The developers of WordNet have paid careful attention to the coherence and completeness of the *IS-A* hierarchy of nouns.

Although earlier versions of WordNet contained several separate *IS-A* hierarchies, the number of separate hierarchies was reduced in successive versions. The top node of each noun hierarchy is called a *unique beginner*. As from WordNet 2.1, the hierarchies have been merged into a single hierarchy headed by the unique beginner *entity*. The noun portion of WordNet may be treated as an ontology of lexicalized concepts. The similarity measures by Resnik [74], Jiang and Conrath [38], Leacock and Chodorow [45], and Lin [48] each exploit the ontology formed by the hierarchy of nouns. To illustrate the WordNet noun hierarchy, a small part of the network surrounding concepts relating to wheeled vehicles is reproduced in Figure 2.8. In this figure solid lines represent *IS-A / SUBSUMES* relations, dashed lines represent *HAS-A / PART-OF* relations and dotted lines represent a series of omitted *IS-A / SUBSUMES* relations.

2.5 Semantic Relatedness Measures

There are two general approaches taken by the different relatedness measures that we will describe. The first approach relies on an examination of the shortest path in the WordNet graph that connects two synsets. This approach is represented by the measures of [94], and Hirst and St-Onge [32]. The second approach exploits the definitions provided

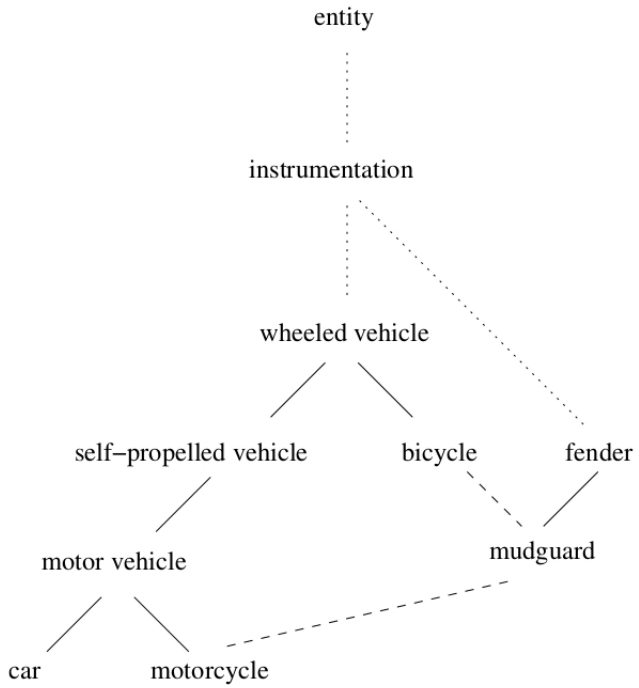

Figure 2.8: *Fragment of the WordNet graph for wheeled vehicles and related concepts.*

for synsets in WordNet, called *glosses*, and is represented by the measure from Banerjee and Pedersen [2].

Sussna

Sussna described one of the first WordNet-based relatedness measures [94]. The measure was developed for the purpose of word sense disambiguation in an information retrieval system. Sussna's measure determines the strength of relatedness between two concepts by first finding the shortest path between their corresponding synsets in the WordNet graph. The edges (the semantic relations) in the path have been assigned with weights, with higher weight indicating greater semantic distance, and the sum of these weights gives the total semantic distance between the concepts.

For example, to compute the relatedness of the concepts *bicycle* and *motorcycle* using Figure 2.8, we would first find the shortest path between these nodes. In this case the path would be:

bicycle HAS-A mudguard PART-OF motorcycle

The semantic distance between *bicycle* and *motorcycle* would therefore be the sum of the distances between *bicycle* and *mudguard*, and between *mudguard* and *motorcycle*. The technique of using the sum of distances on the shortest path between concepts is repeated in many other similarity and relatedness measures, and we will refer to these types of measures as *edge-based measures* (also called path-based measures [82]).

A central problem for path-based measures is determining the distances represented by particular semantic relations in the semantic network. Sussna proposed two schemes for estimating the semantic distances (the "weights") of individual edges in WordNet. He observed that the more concepts a given concept is related to, the less strongly it is associated with each one. More specifically, the semantic distance of a relation is proportional to the number of other semantic relations of the same type emerging from a concept. Sussna calls this the *type-specific fanout* (TSF) factor. For example, the concept for *computer* in WordNet has 14 meronym (HAS-A) relations, corresponding to 14 different parts of a *computer*, such as *keyboard*. The synset including *keyboard*, on the other hand, has only two meronym relations, one of which is *key*. Since *keyboard* has fewer parts than *computer*, *keyboard* will be more strongly associated with each of its parts. Sussna's measure would therefore assign a greater semantic distance value to the meronym link connecting *computer* and *keyboard* than to that connecting *keyboard* and *key*.

The second edge-weighting scheme in Sussna's measure is called *depth-relative scaling*, and is based on the observation that siblings deep in the taxonomy tend to be more closely related than those closer to the top. General, abstract concepts are assumed to represent broad distinctions, and therefore the differences between them cover greater semantic distance than do the finer distinctions found lower in the taxonomy.

To calculate the strength of relatedness between concepts in Sussna's measure, each relationship type is assigned a weight range between min_r and max_r, for each relationship type r. The semantic distance value for a relation of type r from the source node c_1 is:

$$wt(c_1 \rightarrow_r) = max_r - \frac{max_r - min_r}{edges_r(c_1)} \tag{2.13}$$

where $edges_r(c_1)$ is the number of relations of type r originating from c_1. For the hypernymy, hyponymy, holonymy, and meronymy relationships the values min_r and max_r are one and two, respectively. Antonymy links always have a weight of 2.5.

For the purpose of determining the weight of an edge in the path, each edge is assumed to consist of two inverse relations. For example, if *robin IS-A bird*, then it is also the case that *bird SUBSUMES robin*. However, it is possible for the inverse relations to be assigned a different weight by Equation 2.13. For example, the weight for *keyboard HAS-A key* is not necessarily the same as for *key PART-OF keyboard* as we cannot assume that the number of meronyms of *keyboard* and the number of holonyms of *key* are the same. Sussna assumed that the semantic distance between concepts should be a symmetrical relationship and so takes the average of the two weights.

The semantic distance weight of an edge is also scaled by the depth of the relation in the taxonomy. The final semantic distance value for the edge between two adjacent synsets c_1 and c_2 is given by:

$$dist_S(c_1, c_2) = \frac{wt(c_1 \rightarrow_r) + wt(c_2 \rightarrow_{r'})}{2 \times \max(depth(c_1), depth(c_2))} \tag{2.14}$$

In the preceding equation, r is the type of relation that holds between c_1 and c_2, and r' is the inverse of r (the type of relation that holds between c_2 and c_1). To determine the semantic distance between any pair of synsets, Sussna takes the sum of the distances between the nodes in the shortest path between the synsets in WordNet.

Hirst and St-Onge

Hirst and St-Onge proposed a semantic relatedness measure for WordNet that was an adaptation of an earlier measure by Morris and Hirst [32]. The measure was previously based on Roget's thesaurus. Their measure was developed in the context of a system for the automatic detection and correction of malapropisms using lexical chains. Hirst

and St-Onge define a malapropism as "the confounding of an intended word with another word of similar sound or spelling that has a quite different and malapropos meaning." For example, accidentally substituting the word "prostate" for "prostrate" would result in a malapropism.

For their measure, Hirst and St-Onge defined three categories of WordNet relationship types: "upward," "downward" and "horizontal." For example, hypernymy (*IS-A*) is classified as an upward link, as it leads toward the root of the WordNet taxonomy, whereas hyponymy (*SUBSUMES*) is a downward link. In general, the up and down categories are used to separate inverse relations, whereas horizontal link types correspond to relations that do not have inverses. The complete list of classifications used by Hirst and St-Onge is given in Table 2.4.

Table 2.4: *Hirst and St-Onge's classification of WordNet relations into directions.*

Relations	Direction
Also see	Horizontal
Antonymy	Horizontal
Attribute	Horizontal
Cause	Down
Entailment	Down
Holonymy	Down
Hypernymy	Up
Hyponymy	Down
Meronymy	Up
Pertinence	Horizontal
Similarity	Horizontal

Hirst and St-Onge distinguish two strengths of semantic relations: strong and medium-strong. Two words, w_1 and w_2, are strongly related if one of three conditions holds:

1. They are synonyms (there is a synset with both w_1 and w_2).

2. They are antonyms (w_1 and w_2 belong to the synsets c_1 and c_2, and c_1 and c_2 are related by antonymy).

3. One is a compound word that includes the other one, and there exists a semantic relation (of any kind) between synsets containing the words. For example, *school*

and *private school* are strongly related, because *private school IS-A school*, and the compound word *private school* contains *school*.

Medium-strong relations hold between words that have corresponding synsets that are connected in the WordNet graph by an allowable path. A path is allowable if it conforms to one of eight patterns, which are defined in terms of the three directions of semantic links. The motivation for these patterns is the observation that changes in direction often result in reduced overall relatedness. For example, some semantic relations may be viewed as transitive. If A *IS-A* B *IS-A* C, then A *IS-A* C. Similarly, if A *PART-OF* B *PART-OF* C, then A *PART-OF* C. However, when a path includes a change in direction, the transitivity of the relations is compromised. The eight allowable patterns are shown in Figure 2.9. It should be noted that each vector in the patterns in Figure 2.9 represents any number of links in the given direction.

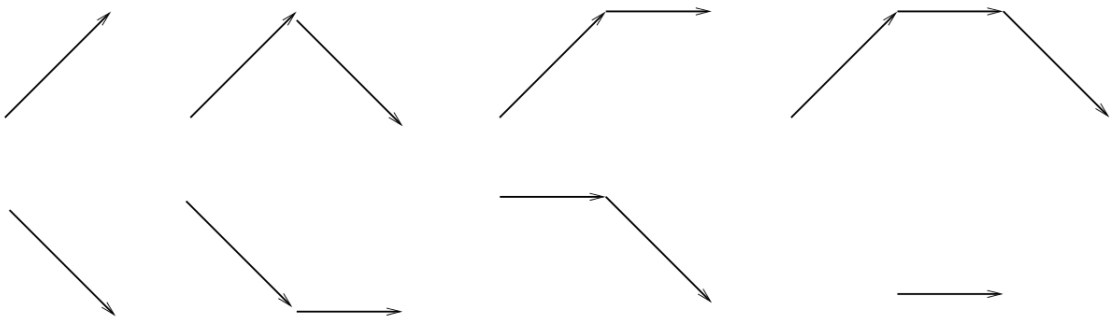

Figure 2.9: *Patterns of semantic relations allowed in medium-strong relationships for Hirst and St-Onge relatedness measure.*

Unlike strong relations, medium-strong relations have a range of relatedness values. The strength of relatedness for a medium-strong relation between the concepts c_1 and c_2 is given by:

$$rel_{HS}(c_1, c_2) = C - len(c_1, c_2) - k \times turns(c_1, c_2) \tag{2.15}$$

where C and k are constants, $length(c_1, c_2)$ is the length, measured in nodes, of the shortest allowable path connecting the synsets c_1 and c_2 , and $turns(c_1, c_2)$ is the number

of changes in direction in the shortest allowable path. Budanitsky and Hirst employed the value eight for C and one for k in their evaluation of the measure [10].

Finally, while the measure described above applies to word senses, in the form of synsets, Hirst and St-Onge also required relatedness values for non-disambiguated word forms. The nodes in WordNet correspond to word senses, but most words have multiple meanings. If it is not known which particular sense of a word is the correct one for the context, then the measure cannot be used as described above. To solve this problem, Hirst and St-Onge assume that the relatedness of word forms is equal to that of their most related senses. Where $S(w_i)$ denotes the set of all senses of the word w_i, the relatedness of the words w_1 and w_2 is:

$$rel(w_1, w_2) = \max_{c_1 \in S(w_1), c_2 \in S(w_2)} [rel_{HS}(c_1, c_2)] \tag{2.16}$$

Banerjee and Pedersen

Banerjee and Pedersen adopt an alternative approach to that of edge-based (path-based) measures, based on a technique by Lesk [2]. Rather than examining paths of semantic relations between word senses, as most other measures do, they compare the text of the definitions provided in WordNet for each synset. Relatedness is computed in terms of the overlap of words in these definitions.

The distinguishing feature of WordNet is the organization of concepts into a semantic network. However, WordNet also supplies short definitions, called *glosses*, for each synset such as might be found in a traditional dictionary. For example, in WordNet 3.0 the gloss for the the first sense of the synset apple is "fruit with red or yellow or green skin and sweet to tart crisp whitish flesh." Banerjee and Pedersen calculate relatedness by counting the number of words that co-occur in the glosses of different synsets.

Banerjee and Pedersen note that phrasal overlaps – sequences of words that appear in different glosses are often indicative of a strong relationship between concepts. They therefore assign a higher value to a phrasal overlap of n words than to an overlap of n words that are not in sequence. Specifically, a phrasal overlap of n words is assigned the

value n^2, whereas n shared words that do not belong to a phrasal overlap are assigned the value n. For example, the gloss for *drawing paper* is "paper that is specially prepared for use in drafting" and the gloss for *decal* is "the art of transferring designs from specially prepared paper to a wood or glass or metal surface." As the phrase "specially prepared" appears in both glosses, it contributes a score of $2^2 = 4$. The word "paper" also appears in both glosses, and contributes a score of one, for a total score of five.

Gloss overlap is a technique that could be applied to any dictionary or lexicon with textual definitions. However, Banerjee and Pedersen exploit WordNet by comparing the glosses of not only the target synsets, but also of their nearest neighbors in the semantic network. For each relation type r, they define a function $r(s_1)$ that returns the gloss of the synset related by r to s_1. For example, the function $hypernym(s_1)$ returns the gloss of the hypernym of the synset s_1. If s_1 is connected to more than one synset by the relation type r, then $r(s_1)$ returns the concatenation of the glosses of each related synset. In addition, they also define a function named $gloss(s_1)$ that returns the gloss for the synset s_1.

Banerjee and Pedersen observe that not every relation is equally helpful for determining relatedness, and suggest that different relations may be more or less useful depending on the particular application. They therefore suggest a general formula for calculating relatedness that can use any arbitrary subset of semantic relations. Let $RELPAIRS$ be a set of pairs of gloss functions, as defined above. The pairs indicate which relations will be compared to one another when computing relatedness. In order to maintain the symmetry of the measure, for any pair $(r_1, r_2) \in RELPAIRS$, the set must also contain (r_2, r_1). This constraint ensures that $rel_{BP}(s_1, s_2) = rel_{BP}(s_2, s_1)$. Given the set of pairs $RELPAIRS$, and two synsets s_1 and s_2, relatedness is calculated using the following equation:

$$rel_{BP}(s_1, s_2) = \sum_{\forall (r_1, r_2) \in RELPAIRS} score(r_1(s_1), r_2(s_2)) \tag{2.17}$$

In the equation above, $score(t_1, t_2)$ is a function that returns the overlap score of two strings t_1 and t_2. As an illustration, given the set $RELPAIRS = \{(gloss, gloss), (hype, hype), (hypo, hypo), (hype, gloss), (gloss, hype)\}$, the relatedness function would

be:

$$\begin{aligned} rel_{BP}(s_1, s_2) = \; & score(gloss(s_1), gloss(s_2)) + score(hype(s_1), hype(s_2)) \\ & + score(hypo(s_1), hypo(s_2)) + score(hype(s_1), gloss(s_2)) \\ & + score(gloss(s_1), gloss(s_2)) \end{aligned} \tag{2.18}$$

2.6 Semantic Similarity Measures

Since our main interest is in semantic similarity measures, we will therefore describe several important WordNet-based similarity measures. Jiang and Conrath distinguish between three approaches in the literature to measuring similarity [38]. They call these edge-based (or edge-counting), node-based (or node-counting), and combined or hybrid approaches.

The edge-based approach relies entirely on the *IS-A* hierarchy. These measures compute similarity in terms of the shortest path between the target synsets in the taxonomy. The degree of similarity is determined on the basis of this path, and generally will correspond inversely with the path length. The first application of this technique to WordNet is typically attributed to Rada et al. [72]. The edge-based technique offers a very intuitive representation of similarity. The principal criticism of the edge-based approach is that it is sensitive to the quality of the taxonomy that is employed. In particular, many authors have noted the inconsistent conceptual density of the WordNet graph, and the problems that this introduces for the reliability of edge-based measures. The edge-based method is equivalent to the path-based approach used in many relatedness measures, except that it is applied to the *IS-A* taxonomy exclusively, and ignores other semantic relationship types.

In order to address the criticisms of the edge-based measures some author have preferred to use taxonomies to determine the relationships between concepts, but to employ external resources (usually corpus statistics) to calculate the value of similarity. These sorts of measures are called node-based, since they discard information about the edges connecting synsets and focus on a few key nodes, which typically includes the two target nodes and their most specific common subsumer in the taxonomy. Resnik and Lin's

measures will be described as examples of the node-based approach.

Finally, while the node-based approach eliminated certain problems that arose from inconsistencies in the taxonomy, it also ignored much useful information that is contained in the paths between synsets. Jiang and Conrath therefore proposed a measure that calculates similarity using the edges in the shortest path, but also uses corpus statistics in a corrective role. However, in our study we have included Jiang and Conrath measure in the second group (node-based) because of the importance played by the nodes in the similarity computation. Those approaches will be introduced in detail in the next two chapters.

2.6.1 Context in Semantic Similarity Measures

The context plays an important role when measuring the similarity of two concepts. Although this is not the main concern of this study, we would like to introduce some works related to this subject.

Lexical co-occurrence counts from large corpora have been used to construct high-dimensional vector-space models of languages. Distances between word vectors extracted from these models are generally considered to reflect semantic similarity. Implicit in this assumption is that "semantic distance" measurements correspond to human intuitions. The validity of one of such measure, "contextual similarity," calculated from spoken the spoken part of the British National Corpus, was also investigated [52]. McDonald in that work, despite on been a pioneer in the issue, provides support for the role of lexical co-occurrence information in modeling semantic similarity.

Keßler explores the influence of context in existing similarity measurement approaches for the geospatial domain, focusing on whether and how these approaches account for it [40]. Based on these observations, the processing of context during similarity measurement is analyzed, and general implementation issues, especially ease of integration into existing reasoning systems and computability, are discussed. The results of the different analyses are then combined into a generic set of characteristics of context for similarity measurement, with regard to the geospatial domain. This approach is also used by Ke-

bler, where a combination of the SIM-DL theory [36], which measures similarity between concepts represented using description logic, and a context model distinguishing between internal and external context to quantify this impact is used [41].

Janowicz et. al. considered one way to make similarity measures context-aware is by introducing weights for specific characteristics [37]. In this work, they proposed a novel approach to semi-automatically adapt similarity theories to the user's needs and hence make them context-aware. Their methodology is inspired by the process of georeferencing images in which known control points between the image and geographic space are used to compute a suitable transformation. The authors proposed to semi-automatically calibrate weights to compute inter-instance and inter-concept similarities by allowing the user to adjust pre-computed similarity rankings. These known control similarities are then used to reference other similarity values.

A different approach is used by Tsinaraki et. al. [98]. They stated contexts may be defined along different dimensions, such as language (Italian, English, French, ...), domain (Philosophy, Computer Science, Physics, ...), time (Ancient Greece, 20th century, ...) etc. Given a conceptual model $\mathcal{M}$ (aka ontology), a context $\mathcal{C}$ and a query $\mathcal{Q}$ they proposed algorithms for interpreting all the terms of the query with respect to $\mathcal{M}$ and $\mathcal{C}$. They also defined and solved the inverse problem: given a set of concepts $\mathcal{S}$ which are part of the answer to query $\mathcal{Q}$ and a context $\mathcal{C}$, they proposed algorithms for choosing terms for all the concepts in $\mathcal{S}$. To illustrate the framework, they used a case study involving a history ontology whose elements are named differently depending on the time period and language of the query.

While many researchers have contributed to the field of semantic similarity models so far, Dong et. al. considered that most of the models are designed for the semantic network environment [20]. When applying the semantic similarity model within the semantic-rich ontology environment, they proposed to solve the following two issues: (1) do not ignore the context of ontology concepts and (2) do not ignore the context of relations. They presented a solution for the two issues, including a novel ontology conversion process and a context-aware semantic similarity model, by considering the factors of both the context

of concepts and relations, and the ontology structure [20].

2.7 Chapter Summary

Throughout the sections above, we have discussed different issues related to the interdisciplinary domain of research known under the name of ontology matching. We have seen that there are many possible definitions of an ontology, just as there are plenty of application fields, ranging from knowledge integration, semantic interoperability, through natural language processing, problem solving and reasoning, to facilitating business interaction and increasing the efficiency of production.

Ontology matching is defined as the process of finding correspondences between the elements of two or more ontologies, which are assumed to cover the same or similar domains of knowledge, but their components are not explicitly mapped to one another. We have presented a classification of ontology matching techniques and approaches. We have particularly emphasized on one major group of technique, which are directly relevant to the approach developed in this thesis: lexical semantic similarities measures.

We have introduced the main ideas behind the edges-based and the nodes-based semantic similarity measures. However, they suffer from a serious limitation. They mostly focus on the semantic information shared by the compared concepts, i.e., on the common points in the concepts definitions or in the semantic differences but they rarely combine both. In the next chapter, we will examine taxonomic features for their use in upgrading existing node-based semantic similarities measures for word pair comparison.

Chapter 3

A Corpus Independent Information Content Metric

Semantic similarity of words pairs is often represented by the similarity between the concepts associated with the words. Several methods have been developed to compute words semantic similarity. Mostly operating on the taxonomic dictionary WordNet and exploiting its hierarchical structure. Node-based methods compute the similarity between two nodes by associating a weight to each node.

In this chapter we first introduce in detail the node-based semantic similarity measures (Section 3.1). Then we study the traditional metrics for computing each node's weight, their properties, limitations (Section 3.2) and from it we propose and test a new corpus independent information content metric (Section 3.3). Experimental results (Section 3.4) and a general discussion about them (Section 3.5) are also covered.

3.1 Motivation

Node-based methods compute the similarity between two nodes by associating a weight to each node. From the perspective of information theory, this weight represents the *information content (IC)* of a concept. The information content can be considered a measure that quantifies the amount of information a concept expresses. The more specialized a concept is, the heavier its weight will be. The conventional way of measuring the IC of word senses is to combine knowledge of their hierarchical structure from an ontology with statistics on their actual usage in text as derived from a large corpus, doing IC based methods corpus dependent.

The increasing need for better measures has led us to this study in the hope of upgrading existing semantic similarities measures. In particular, we exploits some foundations of the Intrinsic Information Content metric (IIC) approach for developing a novel metric for computing the IC [84]. The metric, completely derived from WordNet without the need for external resources from which statistical data is gathered, is applied in combination with recently developed semantic similarity measures to the words pairs comparison problem [55]. However, before getting deeper into how the information content can be computed let's introduce the semantic similarity measures which will make use of those information content metrics.

3.1.1 Node-based Semantic Similarity Measures

One way to compute the similarity between two nodes in a graph is by associating a weight with each node. Such similarity measures are called *node-based similarity measures*. The node-based similarity measures include the metrics of Resnik [74], Jiang & Conrath [38], Lin [48] and Pirró & Seco [70, 67, 68].

Resnik

Resnik introduced the first similarity measure to combine corpus statistics with a conceptual taxonomy [74]. Resnik's approach has received considerable attention, and

a number of other measures have incorporated his technique. Resnik defines similarity in terms of information theory, and derives the necessary probability information from a corpus of text. The key intuition in Resnik's measure is that for any two concepts, the most specific concept that subsumes them both in the conceptual taxonomy represents the information that the concepts share in common. For example, in Figure 2.8 the most specific common subsumer of *car* and *bicycle* is *wheeled vehicle*. The concept *wheeled vehicle* is assumed to represent the information that is common to both *car* and *bicycle*. Resnik determines similarity by calculating the information content of the shared subsumers. That is, higher information content means that the concepts share more in common, and so are more similar.

First, Resnik defined $P(c)$ as the probability of encountering an instance of a concept c. In order to determine $P(c)$, Resnik relied on frequency information from a text corpus. When counting the instances of concepts in the corpus, any instances of subsumed concepts are also counted as instances of their subsuming concept. For example, any instances of the words for *apple, orange, banana*, etc. also count as instances of *fruit*. The concept *fruit* will necessarily have a higher frequency than any concepts it subsumes, including every concept subsumed by its children, and so on. Therefore, the probabilities of encountering concepts increases monotonically for concepts higher in the taxonomy.

In order to compute the probability function $P(c)$, we must first calculate the number of occurrences of the concept c and the occurrences of all concepts subsumed by c. Where $words(c)$ denotes the set of words that correspond to all of the concepts subsumed by c, the total frequency of c is given by:

$$freq(c) = \sum_{n \in words(c)} count(n) \tag{3.1}$$

The probability of encountering a concept c may be defined as the relative frequency of c, where N is the total number of words observed in the corpus:

$$P(c) = \frac{freq(c)}{N} \tag{3.2}$$

For his experiments, Resnik employed the Brown Corpus of American English [1]. He counted only the nouns in this corpus, and only those nouns that are associated with concepts in WordNet.

According to the axioms of information theory, the information content of a concept c is the negative log of its likelihood: $IC(c) = -\log P(c)$.

As mentioned above, Resnik argued that the similarity of two concepts is proportional to the amount of information that they share, and that the shared information is represented by their most specific common subsumer. For example, the most specific shared subsumer of *car* and *motorcycle* in Figure 2.8 is *motor vehicle*. Therefore *motor vehicle* is assumed to represent all of the information that is common to the concepts *car* and *motorcycle*. The amount of information conveyed by the concept *motor vehicle*, as determined by information theory, corresponds to the degree of similarity between *car* and *motorcycle*.

Formally, where $S(c_1, c_2)$ denotes the set of concepts that subsume both c_1 and c_2, the degree of similarity is:

$$\begin{aligned} Sim_{res}(c_1, c_2) &= \max_{c \in S(c_1,c_2)}[-\log P(c)] \\ &= \max_{c \in S(c_1,c_2)} IC(c) \\ &= IC(lcs(c_1, c_2)) \end{aligned} \tag{3.3}$$

A few features of the preceding formula are worth noting. First, similarity always decreases lower in the taxonomy, as information content correlates inversely with $P(c)$. As the root node of the conceptual hierarchy subsumes every concept, it has a probability of exactly one and therefore has an information content of zero. In other words, knowing that two concepts share the root node as a subsumer provides no information, as this is true of any two concepts. If the only common subsumer of two concepts is the root node, they have the least possible similarity. Second, Resnik's equation uses the common subsumer with the maximum information content. This will always be the most specific, i.e. the "lowest," concept in any sequence of super ordinates in the taxonomy. Budanitsky

[1] The Brown Corpus: A Standard Corpus of Present-Day Edited American English. Published in 1961 and revised and amplified in 1979, http://icame.uib.no/brown/bcm.html

and Hirst reformulated Resnik's measure to explicitly refer to the lowest superordinate in the taxonomy [10]. Although Budanitsky and Hirst's formulation is more intuitive than Resnik's, it introduces ambiguity in cases of multiple inheritance. In these cases, it may not be possible to identify a "lower" subsumer, but the information content gives an indication of the most specific concept.

Jiang and Conrath

Jiang and Conrath proposed a hybrid semantic similarity measure which by convenience we will introduce now among the node-based similarities [38]. Jiang and Conrath sought to combine the advantages of the edge-based and node-based approaches. In order to compensate for the unreliability of edge-distances, Jiang and Conrath weight each edge by associating probabilities based on corpus statistics. Their approach is similar to Resnik's, in that it employs information from both a conceptual taxonomy and from a text corpus. However, whereas Resnik bases the value of similarity on the information content of one node – the most informative common subsumer – Jiang and Conrath use information theory to determine the weight of each link in a path.

Jiang and Conrath argue that the degree of similarity between a parent and its child in the noun hierarchy of WordNet is proportional to the probability of encountering the child, given an instance of the parent: $P(c \mid par(c))$. By definition, the quantity $P(c \mid par(c))$ is:

$$P(c \mid par(c)) = \frac{P(c \bigcap par(c))}{P(par(c))} \tag{3.4}$$

Like Resnik, Jiang and Conrath consider every instance of a child to be an instance of its parent, and thus $P(c \bigcap par(c)) = P(c)$. That is, it is redundant to require both a child c and its parent $par(c)$, as every instance of c is also an instance of $par(c)$. The equation for the probability of a child, given an instance of its parent, can therefore be simplified to:

$$P(c \mid par(c)) = \frac{P(c)}{P(par(c))} \tag{3.5}$$

Jiang and Conrath define the semantic distance between a child c and parent $par(c)$ as the information content of the conditional probability of c given $par(c)$, and using the basic properties of information theory obtain the following semantic distance equation:

$$\begin{aligned} Dist_{j\&c}(c, par(c)) &= -\log P(c \,|\, par(c)) \\ &= IC(c \bigcap par(c)) - IC(par(c)) \\ &= IC(c) - IC(par(c)) \end{aligned} \tag{3.6}$$

The semantic distance between a parent and its child concept is therefore the difference in their information content. This seems a plausible conclusion, as the difference in information content should reflect the information required to distinguish a concept from all of its sibling concepts. For example, if a parent has only a single child, then the conditional probability $P(c \,|\, par(c)) = 1$. In this case, taking the negative logarithm gives $dist_{j\&c} = 0$. If no additional information is required to distinguish a child from its parent, then the semantic distance between them ought to be zero; they are effectively the same concept.

To compute the total semantic distance between any two concepts in the taxonomy, Jiang and Conrath's measure uses the sum of the individual distances between the nodes in the shortest path. As the shared subsumer (denoted by $lcs(c_1, c_2)$ for the lowest common subsumer or lowest super-ordinate shared by c_1 and c_2) does not have a parent in the path, this node is excluded from the summation. The semantic distance between any two concepts c_1 and c_2 in the taxonomy is therefore:

$$dist_{j\&c}(c_1, c_2) = \sum_{c \in path(c_1,c_2) \setminus lcs(c_1,c_2)} dist_{j\&c}(c, par(c)) \tag{3.7}$$

By substituting the expression in Equation 3.6 into Equation 3.7 and expanding the summation, we obtain:

$$Dist_{j\&c}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \tag{3.8}$$

Lin

Lin attempted to provide a more general and theoretically sound basis for determining the similarity between concepts than previous work had provided [48]. He argued that similarity measures should not depend on the domain of application, nor on the details of the resources that they use. Lin begins by proposing three key intuitions about similarity:

- Intuition 1: The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.
- Intuition 2: The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.
- Intuition 3: The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

Lin argued that as there are different ways of capturing the intuitions above, an additional set of assumptions are required. Lin therefore proposed a set of five assumptions that capture these intuitions, and from which a measure of similarity may be derived. The five assumptions are stated in terms of information theory. In the following assumptions, $common(A, B)$ is a proposition that states the commonality of the objects A and B, and $description(A, B)$ is a proposition that states what A and B are.

- Assumption 1: The commonality between A and B is measured by:
  $IC(common(A, B))$
- Assumption 2: The difference between A and B is measured by:
  $IC(description(A, B)) - IC(common(A, B))$
- Assumption 3: The similarity between A and B is a function of the commonalities and differences of A and B. Formally:
  $sim(A, B) = f(IC(common(A, B)), IC(description(A, B)))$
- Assumption 4: The similarity between a pair of identical objects is always one. Thus: $sim(A, A) = 1$

- Assumption 5: The similarity between a pair of objects with no commonality is always zero. Thus: $\forall y > 0, f(0, y) = 0$

- Assumption 6: If the similarity between A and B can be computed using two independent sets of criteria, then the overall similarity is the weighted average of the two similarity values:
  $\forall x_1 \leq y_1, x_2 \leq y_2 : f(x_1 + x_2, y_1 + y_2) = \frac{y_1}{y_1+y_2} f(x_1, y_1) + \frac{y_2}{y_1+y_2} f(x_2, y_2)$

Using the six assumptions listed above, Lin proves the following similarity theorem:

$$sim_{lin}(A, B) = \frac{\log P(common(A, B))}{\log P(description(A, B))} \tag{3.9}$$

In order to apply the similarity theorem above to a conceptual taxonomy, Lin follows similar reasoning to that of Resnik. The concept in a taxonomy that corresponds to the statement of the commonalities between the concepts c_1 and c_2 is the lowest common superset, denoted $lcs(c_1, c_2)$. Similarly, the statement that describes the concepts c_1 and c_2 is the union of the two concepts. The information content of the statement "c_1 and c_2" is the sum of the information content of c_1 and c_2 . According to the basic premise of information theory the information content of a message is the negative log of its probability, and therefore the sum of the information content of c_1 and c_2 is $-\log P(c_1) + -\log P(c_2)$. Substituting into Lin's similarity theorem, we obtain:

$$Sim_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{3.10}$$

Lin's measure is therefore the ratio of the information shared in common to the total amount of information possessed by two concepts. It is quite similar to Resnik's measure except that Resnik's measure considers only the information that is shared by concepts, and does not take into account the total amount of information that they represent. Due to this, Resnik's measure cannot distinguish between different pairs of concepts that have the same most informative subsumer. For example, in the small semantic network in Figure 2.8, the concept pair *car/bicycle* has exactly the same similarity as the pairs motor *vehicle/bicycle* and *self-propelled vehicle/bicycle* according to Resnik's measure.

Pirró and Seco

The Pirró and Seco similarity metric is based on Tversky's theory [99] but from an information-theoretic perspective [67, 70]. This measure achieves very good results in the comparison to human judgments when it is combined with the notion of *intrinsic information content*. In the next section different approaches for computing the information content (IC) will be introduced.

$$Sim_{p\&s}(c_1, c_2) = \begin{cases} 3IC(lcs(c_1, c_2)) - IC(c_1) - IC(c_2) & if\ c_1 \neq c_2 \\ 1 & if\ c_1 = c_2 \end{cases} \tag{3.11}$$

Pirró and Euzenat proposed a modification of the work in [67, 70] for extending the similarity introduced in Equation 3.11, so it can also consider relations beyond inheritance by changing the way in which the information content was computed [68]. They also presented the new semantic relatedness by using the contrast model perspective, but this last topic will be covered in the next chapter.

3.2 Information Content Metrics

Information content can be considered a measure that quantifies the amount of information a concept expresses. The more specialized a concept is, the higher the amount of information it should carry out. The literature describes two main approaches of computing the information content:*corpus dependent* and *corpus independent*.

3.2.1 Corpus Based Information Content Metric

As a way of eliminating the unreliability of edges distances in a taxonomy when none or more than one path exist, Resnik proposed to associate probabilities to the concepts in the taxonomy [74]. Resnik proposed to augment the taxonomy with a function $P : C \rightarrow [0, 1]$, such that for any $c \in C$, $P(c)$ is the probability of encountering an instance of concept c in a corpus. This implies that P is monotonic as one moves up in the taxonomy: if c_1

IS-A c_2 , then $P(c_1) \leq P(c_2)$. Moreover, if the taxonomy has a unique top node then its probability is 1.

Resnik followed the standard argumentation of information theory, the information content of a concept c can be quantified as negative the log likelihood, see Equation 3.12. Quantifying the information content in this way makes intuitive sense in this setting: as probability increases, informativeness decreases, so the more abstract a concept, the lower its information content. Moreover, if there is a unique top concept, its information content is 0, $IC(top\_concept) = 0$.

$$IC(c) = -\log\ P(c) \tag{3.12}$$

3.2.2 Corpus Independent Information Content Metrics

Intrinsic Information Content Metric

In previous section we presented Resnik's approach of computing information content [74]. However in this approach the information content is obtained through statistical analysis of copora, from where probabilities of concepts occurring are inferred. Seco et. al. probed WordNet can also be used as a statistical resource with no need for external ones [67, 70, 84]. Moreover, they argue that the WordNet taxonomy may be innovative exploited to produce the IC values needed for semantic similarity calculations.

Their method of obtaining IC values rests on the assumption that the taxonomic structure of WordNet is organized in a meaningful and principled way, where concepts with many hyponyms convey less information than concepts that are leaves. They argued that the more hyponyms a concept has the less information it expresses, otherwise there would be no need to further differentiate it. Likewise, concepts that are leaf nodes are the most specified in the taxonomy so the information they express is maximal. Hence, they express the information content values of a WordNet concept c as a function of the hyponyms it has. Formally shown in Equation 3.13:

$$IIC(c) = \frac{\log\left(\frac{hypo(c)+1}{\max_{wn}}\right)}{\log\left(\frac{1}{\max_{wn}}\right)} = 1 - \frac{\log\left(hypo(c)+1\right)}{\log\left(\max_{wn}\right)} \tag{3.13}$$

where the function $hypo(c)$ returns the number of hyponyms of a given concept c and max_{wn} is a constant that is set to the maximum number of concepts that exist in the taxonomy.

The denominator, which is equivalent to the value of the most informative concept, serves as a normalizing factor in that it assures that IC values are in $[0, 1]$. The above formulation guarantees that the information content decreases monotonically. Moreover, the information content of the imaginary top node of WordNet would yield an information content value of 0.

This metric was extended to consider multi *parts-of-speech* of the WordNet taxonomy rather than just the noun taxonomy [69, 68]. The authors extend the idea of IIC to adjectives and adverbs by taking into account their relations with nouns and verbs. They stand that, adjectives and adverbs are related to nouns and verbs by semantic relations enabling to assess features of each synset in terms of IC. In particular, for each adjective and adverb synset, the multi part-of-speech IC (IC_m) for each synset S is defined as follows:

$$IIC_m(c) = \sum_{j=1}^{m} \frac{\sum_{k=1}^{n} IIC(c_k \in C_{R_j})}{|C_{R_j}|} \tag{3.14}$$

This formula takes into account all the m kinds of relations that connect a given adjective or adverb synset S with nouns and verbs. In particular, for all the synsets at the other end of a particular relation (i.e., each $c_k \in C_{R_j}$) the average IIC is computed. This enables to take into account the expressiveness of an adjective or adverb in terms of its relations with nouns and verbs.

However the authors also used the same approach to consider relations beyond inheritance [68]. They evaluate each type of ontological relation between concepts to provide a better indicator about the features of concepts which can be used to compute relatedness. For instance, by only focusing on *IS-A* relations, in the Figure 3.1 we would lose some important information (i.e., that *car* has *part-of engine* or that *bicycle* has a *part-of sprocket*) that can help to further characterize commonalities and differences between the two concepts.

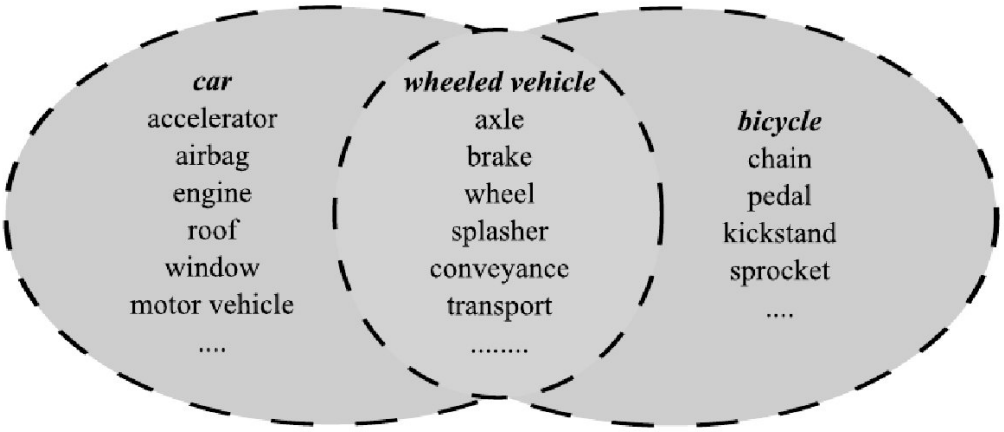

Figure 3.1: *An example of concepts features.*

$$EIC(c) = \sum_{j=1}^{m} \frac{\sum_{k=1}^{n} IIC(c_k \in C_{R_j})}{|C_{R_j}|} \tag{3.15}$$

This formula takes into account all the m kinds of relations that connect a given concept c with other concepts. Moreover, for all the concepts at the end of a particular relation (i.e., each $c_k \in C_{R_j}$) the average IIC is computed. This enables to take into account the expressiveness of concepts to which a concept is related in terms of their information content. The final value of *Extended Information Content* (eIC) is computed by weighting the contribution of the IIC and EIC coefficients:

$$eIC(c) = \zeta IIC(c) + \eta EIC(c) \tag{3.16}$$

The two parameters ζ and η can be settled in order to give more or less emphasis to the hierarchical IC of the two concepts.

Saruladha et. al. proposed an extension to the intrinsic information content IIC. Saruladha's approach to assess semantic similarity among concepts from different and independent ontologies without constructing a priori a shared ontology. They said finding the *lcs* of concepts belonging to different ontologies is possible when both ontologies are connected through a virtual root [81].

Recently, Cross and Yu proposed a different approach where a fuzzy set-theoretic interpretation is given to the features selected to describe a concept in an ontology [15]. They criticized the extended Information Content metric (eIC) [68] for being a parameterized weighting of the intrinsic information content (IIC) and its total average relationship (EIC). Because $EIC(c)$ is the summation for each kind of relationship k, of the average of the $IIC(c_i)$ for all concepts c_i that are "at the end of a particular relation" [68], i.e., k, with concept c. Cross and Yu said that how non-taxonomic relationships and their inverses are handled, i.e., how is "at the other end of a particular relation" was not clear. They stand that if inverse relationships are not ignored or inverse relationships are not clearly identified, a circular calculation of eIC could occur. They also were concerned about how and which non-taxonomic relationships were used [15].

Ontology-Based Information Content Computation

Even though intrinsic IC computation has led to accurate assessments, Sánchez et. al though that there was still room for improvement [79, 78]. They stand that, in an ontology, taxonomic leaves represent the semantic of the most specific concepts of a domain. So, the set of leaves belonging to a domain would accurately define its scope.

Sánchez et. al. criticized the work of Zhou [106] because of the inclusion of a parameter which should be empirically tuned [79]. However they agreed the depth was an important dimension in order to differentiate degrees of concreteness. Sánchez et. al. said the depth of a concept in a taxonomy corresponds in fat to its number of taxonomic subsumers (when no multiple inheritance is considered). So, the larger the amount of subsumer

concepts above a given one, the higher its degree of concreteness as it is the result of many specializations [79].

Finally defining the IC metric as follows:

$$IC_{ONT}(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \tag{3.17}$$

where max_leaves represents the maximum number of leaves in the ontology, $subsumers(c)$ stands for the concepts from which c is an specialization of and $leaves(c)$ was defined as follows:

Definition 3.2.1. ***(Leaves)*** *Let $\mathcal{C}$ be the set of concepts of the ontology then the set of leaves of a concept c is defined as:*

$leaves(c) = \{l \in \mathcal{C} | \, l \in \texttt{hyponyms(c)} \wedge \texttt{l is a leaf}\}$

where l is a leave iff $hyponyms(l) = \emptyset$.

Web-Based Information Content Metric

In other work Sánchez et. al. decided to exploit the Web as a corpus [80] . Under the assumption the Web as a social-scale general purpose corpus. They said, its main advantages were its free and direct access and its wide coverage of any possible domain. In comparison with other general purpose repositories (such as the Brown Corpus) which have shown a poor performance for domain-dependent problems, the Web's size is millions of orders of magnitude higher. In fact, the Web offers more than 1 trillion of accessible resources which are directly indexed by web search engines [80].

For using the Web as a corpus, the authors compute term occurrences from the Web instead of a reliable, closed and domain-specific repository. The main problem of computing term's Web occurrences is that the analysis of such an enormous repository for computing appearance frequencies is impracticable. However, the availability of massive Web Information Retrieval tools (general-purpose search engines like Google) can help in this purpose, because they provide the number of pages (*hits*) in which the searched terms occur.

This is that the probabilities of Web search engine terms, conceived as the frequencies of page counts returned by the search engine divided by the number of indexed pages, approximate the relative frequencies of those searched terms as actually used in society. So, exploiting Web Information Retrieval (IR) tools and concept's usage at a social scale as an indication of its generality, one can estimate, in an unsupervised fashion, the concept probabilities from Web hit counts.

Even though web-based statistical analyses brought benefits to domain-independent unsupervised approaches (i.e. no background ontology is exploited), due to their lack of semantics, their performance is still far from the other supervised (ontology-based) measures. Taking those aspects into consideration, Sánchez et. al. adapted the IC-based similarity measures to exploit the Web as a corpus, by estimating concept's IC from web hit counts [80]. However, the *LCS* is extracted from the ontology. The Web-based IC computation is specified as follows:

$$IC_{IR}(c) = -\log p_{web}(c) = -\log \frac{hits(c)}{total\_webs} \tag{3.18}$$

Where, $p_{web}(c)$ the probability of appearance of string 'c' in a web resource. This probability is estimated from the Web hit counts returned by Web Information Retrieval tool -*hits*- when querying the term 'c', *total_webs* is the total number of resources indexed by a web search engine.

In the next section we present the limitations of the corpus based approach and the intrinsic information content approach. From it we propose a new metric for computing the information content.

3.3 Extending the Intrinsic Information Content

As pointed out before, semantic similarity measures grounded on information content obtain IC's values for concepts by statistically analyzing large corpus and associating a probability to each concept in the taxonomy based on its occurrences within the considered corpus. From a practical point of view, this approach has two main drawbacks:

1. it is time consuming, and

2. it heavily depends on the type of corpus considered.

Toward mitigating these drawbacks Seco et al. proposed the Intrinsic Information Content (IIC), see Equation 3.13 [84]. The information content values of concepts, rest on the assumption that the taxonomic structure of WordNet is organized in a "meaningful and structured way," where concepts with many hyponyms convey less information than concepts that are leaves, that is, the more hyponyms a concept has the less information it expresses.

From Seco et al. perspective, concepts that are leaf nodes are the most specific in the taxonomy so the information they express is maximal. This means it does not matter how deep is the leaf in the taxonomy. However, our approach, although founded in the same ideas, is different.

Sustained in the notion, a new concept is created when none of the existing concepts is unable to describe all the properties or characteristics of the new phenomenon or object; and the new concept goes to enrich the knowledge base of humanity. If a taxonomy's depth is smaller than another taxonomy's depth then the first taxonomy enclose less amount of knowledge than the second one. We stand the depth in the taxonomy it is also an important factor to consider. *The deeper a concept is found in a taxonomy means the amount of previous knowledge is larger and it should bear a higher value of information content.*

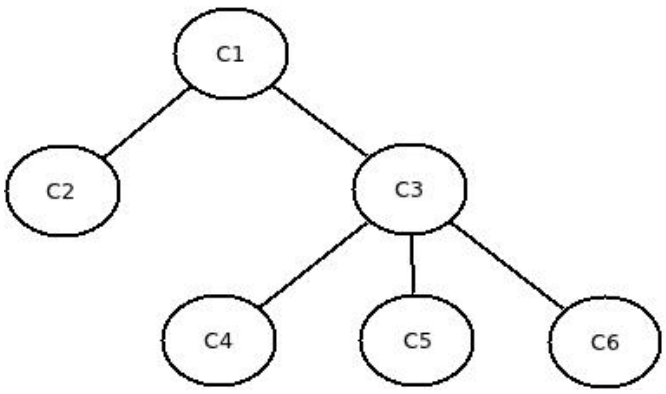

Figure 3.2: *Abstract taxonomy.*

For example, in Figure 3.2 concepts c_2 and c_4 should not have the same information content value, but since they both have the same number of hyponyms under Seco et al. approach they have equal values of information content.

However, considering Figure 3.2 we say the similarity between concepts c_2 and c_1 should be greater than the similarity between concepts c_4 and c_1: $(Sim(c_2, c_1) > Sim(c_4, c_1))$, but if we use Seco et al. intrinsic information content approach (IIC) [84] this is not true:

$$\begin{aligned} hypo(c_2) &= hypo(c_4) \\ IIC(c_2) &= IIC(c_4) \\ 3IIC(c_1) - IIC(c_2) - IIC(c_1) &= 3IIC(c_1) - IIC(c_4) - IIC(c_1) \\ substituting\ 3.11 & \\ Sim_{p\&s}(c_2, c_1) &= Sim_{p\&s}(c_4, c_1) \end{aligned} \tag{3.19}$$

We have to clarify that, as well as generally assumed in WordNet, we are considering each of the edges hold the same amount of importance, which indeed it is not true. In the next subsection we introduce the mathematical foundation behind our idea.

To avoid the previous situation there are some properties which should be fulfilled by our IC metric:

$$IC(c) \propto depth(c)$$

$$IC(c) \propto \tfrac{1}{hypo(c)}$$

$$\begin{aligned} IC(c_1) = IC(c_2) \Leftrightarrow\ & hypo(c_1) = hypo(c_2) \\ & \wedge depth(c_1) = depth(c_2) \\ IC(c_1) < IC(c_2) \Leftrightarrow\ & hypo(c_1) > hypo(c_2) \\ & \wedge depth(c_1) = depth(c_2) \\ IC(c_1) < IC(c_2) \Leftrightarrow\ & hypo(c_1) = hypo(c_2) \\ & \wedge depth(c_1) < depth(c_2) \end{aligned} \tag{3.20}$$

Then, our corpus independent information content metric will be a function of the *number of hyponyms* ($hypo(c)$) and the *concept's depth* ($depth(c)$). Considering the above properties we proposed the following expression for the new information content metrics, see Equation 3.21.

$$IC_{hd}(c) = 1 - \left(\frac{\log\left((hypo(c)+1) * (max_{depth} - depth(c)+1)\right)}{\log\left(max_{wn} * max_{depth}\right)} \right) \tag{3.21}$$

where function $hypo(c)$ returns the number of hyponyms of a given concept c, max_{wn} is a constant that is set to the maximum number of concepts that exist in the taxonomy, function $depth(c)$ returns the depth of a given concept c in the taxonomy and max_{depth} represents the maximum depth of the corresponding taxonomy.

The denominator, which is equivalent to the value of the most informative concept, serves as a normalizing factor in that it assures that information content values are in the range $[0, 1]$. The above formulation guarantees that the information content decreases monotonically. Moreover, the information content of the imaginary top node of WordNet would yield an information content value of 0.

When the number of hyponyms of a concept ($hypo(c)$) decrease, the fraction in the IC_{hd} expression tends to 0 and it moves IC_{hd} metric closer to its maximum value, 1. Similar behavior is observed when the concept is located deeper into the taxonomy. The difference between the maximum depth of the corresponding taxonomy and the concept's depth $depth(c)$ moves closer to 0. When this difference is closer to 0 the fraction in the IC_{hd} expression tends to 0 and it moves IC_{hd} metric closer to its maximum value, 1.

3.4 Experiments and Results

3.4.1 Experimental Settings

The purpose of the experiment was to evaluate the new information content metric. We use it into some information content based similarity measures: Lin's measure (Equation 3.10), Resnik's measure (Equation 3.3) and Pirró & Seco's measure (Equation 3.11)

and compare the results through the performance of those similarity measures with and without using our new information content metric.

As a baseline for subsequent comparisons we started with the results presented in [55]. We used the human judgments of Pirró and Seco experiment (P&S in the following) for the word pairs on the Miller and Charles dataset (M&C in the following) and the Rubenstein and Goodenough dataset (R&G in the following) [70].

Unfortunately, there is a distinct lack of standards for evaluating semantic similarities, which means that the accuracy of a computational method for evaluating words similarity can only be established by comparing its results against human common sense. That is, a method that comes close to matching human judgments can be deemed accurate.

The Pearson correlation coefficient indicates the strength of a linear relationship between two variables. Although its value generally does not completely characterize their relationship, we will use it for comparing the results of our similarity measures and the human judgments. The Pearson correlation is $+1$ in the case of a perfect positive (increasing) linear relationship, -1 in the case of a perfect decreasing (negative) linear relationship, and some value between -1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches to zero there is less of the relationship. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables. If the variables are independent, Pearson's correlation coefficient is 0, but the inverse is not true because the correlation coefficient detects only linear dependencies between two variables.

Some datasets of word pairs are commonly used for this evaluation. In particular, the Miller and Charles dataset (M&C in the following) and the Rubenstein and Goodenough dataset (R&G in the following) are standard datasets for evaluating semantic similarities, see Table A.1 and Table A.2 in Appendix A.

In 1965, Rubenstein and Goodenough obtained "synonymy judgments" of word pairs by hiring 51 subjects to evaluate 65 pairs of nouns [76]. The subjects were asked to assign a similarity from 0 to 4, from "semantically unrelated" to "highly synonymous." Miller and Charles, 25 years later, extracted 30 pairs of nouns from the R&G dataset

and repeated their experiment with 38 subjects [58]. The M&C experiment achieved a correlation of 0.97 with the original experiment of R&G. Resnik, in 1995, replicated the M&C experiment with 10 computer science students, obtaining a correlation of 0.96 [74]. Pirró and Seco in 2008 also recreated the R&G experiment this time with 101 subjects, and arrived at a correlation coefficient of 0.972 for the full dataset [70].

We used the human judgments of Pirró and Seco experiment for the word pairs of both datasets, the Miller and Charles dataset (M&C) and the Rubenstein and Goodenough dataset (R&G), see Table A.1 and Table A.2 in Appendix A [70]. In M&C dataset we considered only 28 word pairs of the 30 used in the M&C experiment since a word missing in WordNet 3.0 made it impossible to compute ratings for the other two word pairs.

All the evaluations were performed using WordNet 3.0 [24] and the Brown Corpus[2] was used for the calculation of the corpus-dependent information content metric. For the computation we used Pedersen's WordNet::Similarity Perl module as the core. We also recreated the Pirró and Seco's experiment with the Java WordNet Similarity Library (JWSL) [70] using Pirró and Seco's intrinsic information content (IIC), but we did not obtain the same results they did. Probably due to the selection of the parameters. Table 3.1 shows the values we used for the parameters during the computation of the corpus independent information content metric (IIC and IC_{hd}) for the nouns and verbs WordNet's sub-taxonomies.

Table 3.1: *Parameters used for corpus-independent information content computation.*

Parameter	Taxonomy	Value
$\max_{wn}$	Noun	82115
	Verb	25047
$\max_{depth}$	Noun	20
	Verb	14

Table 3.2 shows some examples of the information content value for various concepts using different IC computation approaches. The columns represent: the concept's string, the taxonomy where is located (noun, verb, adj, adv), the corresponding sense's number, the total number of hyponyms the concept have in the taxonomy, depth of the concept,

[2]The Brown University Standard Corpus of Present-Day American English.

the IC value (corpus-dependent), the IIC value (Seco's approach) and the IC_{hd} value (our approach).

Table 3.2: *Values of the information content for various concepts using different information content approaches.*

concept	pos	sense	hypo	depth	IC	IIC	IC_{hd}
entity	noun	1	74373	2	0.0	0.009	0.018
cock	noun	4	1	15	12.16	0.939	0.875
noon	noun	1	0	11	11.06	1.0	1.0

3.4.2 Experimental Results

Table 3.3 is a summary of the evaluations of the proposed metric (IC_{hd}). It compiles the correlation values of node-based similarity measures (Sim_{lin} (Equation 3.10), Sim_{res} (Equation 3.3), $Sim_{j\&c}$ (Equation 3.8) and $Sim_{p\&s}$ (Equation 3.11)) when compared with the human judgment of P&S experiment for the traditional corpus based information content, the intrinsic information content and our proposed metric (IC, IIC and IC_{hd} respectively). The evaluation was done using the 28 word pairs M&C dataset and the 65 word pairs R&G dataset.

Table 3.3: *Maximum values of correlation obtained for node-based semantic similarity measures using different IC metrics for the M&C and the R&G word pairs datasets.*

		M&C	R&G
Sim_{lin}	IC	0.8587	0.8812
	IIC	0.8797	0.8992
	IC_{hd}	**0.8821**	**0.9007**
Sim_{res}	IC	0.8308	0.8672
	IIC	**0.8421**	**0.8773**
	IC_{hd}	0.8361	0.8679
$Sim_{j\&c}$	IC	-0.8660	-0.8689
	IIC	**-0.8805**	**-0.8848**
	IC_{hd}	-0.8712	-0.8747
$Sim_{p\&s}$	IC	0.8655	0.8793
	IIC	**0.8843**	**0.8944**
	IC_{hd}	0.8835	0.8915

3.4.3 Significance Analysis of the Results

To asses the quality of the results a significance test was done for the obtained correlations. Since the dataset was the same (M&C or R&G) we choose the "difference between two dependent correlations test." With this test we can see if the correlation obtained using our new information content metric (IC_{hd}) is significantly different than the correlation obtained using the (IC) and the (IIC) information content metrics.

Since the difference between the means of samples from two normal distributions is itself distributed normally, the T-distribution can be used to examine whether that difference can reasonably be supposed to be zero. The difference between two dependent correlations test assumes as the *null hypothesis* there is correlation between the tested correlations, despite the redundancy. While the *alternative hypothesis* is they are not correlated.

Table A.3, included in Appendix A, contains the upper critical values of the Student's T-distribution. The most commonly used significance level is $\alpha = 0.05$. For a two-sided test, we computed the percent point function at $\alpha/2$ (0.025). If the absolute value of the test statistic is greater than the upper critical value, then we reject the null hypothesis.

The expression for computing the t-values for the difference of correlations between datasets X_1 and X_2 is as described in Equation 3.22:

$$t_{X_1X_2} = \frac{(r_{YX_1} - r_{YX_2})\sqrt{(n-3)(1+r_{X_1X_2})}}{\sqrt{2(1 - r_{YX_1}^2 - r_{YX_2}^2 - r_{X_1X_2}^2 + 2r_{YX_1}r_{YX_2}r_{X_1X_2})}} \tag{3.22}$$

r_{AB} : `stands for the correlation value between datasets` A `and` B.

Returning to our experiment, Table B.1, included in Appendix B, show the results of computing Lin's similarity (Sim_{lin}) for each word pair in M&C dataset as well as some statistics about the obtained results. A more detailed analysis of the correlations values shown in Table 3.3 for Sim_{lin} is shown in Table 3.4.

Table 3.5 shows the results of the significance tests for Lin's similarity for a sample size (N) of 28 and 65. When those results are compare with the upper critical values

Table 3.4: *Correlation values for Sim_{lin} in the M&C dataset using different IC metrics.*

	Sim_{lin}		
Correlation values	**LIN -** IC	**LIN -** IIC	**LIN -** IC_{hd}
P&S ratings	0.8587	0.8797	0.8821
LIN - IC	-	0.9893	0.9892
LIN - IIC	-	-	0.9982

of Student's T-distribution for (N-3) degrees of freedom, see Table A.3, we arrive to the following conclusions about the significance of the obtained correlations for Lin's similarity Sim_{lin}:

Table 3.5: *Significance values for Sim_{lin} in the M&C dataset using different information content metrics.*

Values of the t-statistics for Sim_{lin}			
Sample size	IC **vs.** IIC	IC **vs.** IC_{hd}	IIC **vs.** IC_{hd}
28	1.5291	1.7369	0.4520
65	2.4080	2.7353	0.7119

- The combination of Sim_{lin} with the information content IC_{hd} was significantly different than when the traditional corpus-based IC metric was used for both sample size of 28 and 65.

- The combination of Sim_{lin} with the information content IC_{hd} was not significantly different than when the intrinsic information content IIC metric was used for a sample size of 28 neither for a sample size of 65.

- The combination of Sim_{lin} with the IIC metric was not significantly different than when the traditional corpus-based IC metric was used for a sample size of 28.

- However for a sample size of 65 the results of the combination of Sim_{lin} with both IIC and IC_{hd} information content were significantly different than when the traditional corpus-based IC metric was used.

Table 3.6 shows a descriptive analysis of the obtained results for Lin's similarity (Sim_{lin}) using different information content metrics. Those results showed although the median for the IC_{hd} metric was a little greater than IC and IIC metrics, the standard deviation and the standard error were lower than for those metrics.

Table 3.6: *Descriptive analysis of Sim_{lin} using different IC metrics.*

	Sim_{lin}			
	P&S ratings	**LIN -** IC	**LIN -** IIC	**LIN -** IC_{hd}
Min	0.4211	0.0000	0.0095	0.0167
Max	3.4210	1.0000	1.0000	1.0000
Average	1.7342	0.4971	0.4952	0.4888
Median	1.5928	0.3071	0.3153	0.3289
Mode	-	1.0000	1.0000	1.0000
STDEV	1.0159	0.3857	0.3854	0.3741
STDER	-	0.5306	0.4923	0.4874

The same analysis of significance was done for the other node-based similarity measures Sim_{res}, $Sim_{j\&c}$ and $Sim_{p\&s}$ when using different IC metrics, as shown in Table B.2, Table B.3, Table B.4, Table B.5, Table B.6, Table B.7, Table B.8, Table B.9, Table B.10, Table B.11, Table B.12 and Table B.13, in Appendix B. The results shown in there confirmed our previous results about the competitiveness and stability of the new developed information content metric IC_{hd}.

3.4.4 Further Experimentation

With the goal of narrowing our understanding about for what kind of data our method works well and for what kind of data it does not work that well. We decide to do some further experimentation. This experimentation can be divided into two stages:

1. prediction of the features and their behavior which better characterize the sample data,

2. a detailed ranking analysis of the similarity results in correspondence with the identified features.

Predicting the attributes which better characterize the data

During this experiment we predicted several models for characterizing the human judgment of the 101 participants in the P&S experiment by identifying the attribute or the set of attributes which better do the task. In the experiment we also asses the importance of each attribute in the characterization.

Supported by the Weka Data Mining Software [30] a least median squared linear regression was predicted for a semantic similarity measure (Equation 4.17) when using our model of information content computation (IC_{hd}) and when using the Seco's approach (IIC). The linear regression was done using the human judgments and using a cross-validation. Because of the small size of the dataset (65 examples) for training the classifier, the obtained model was cross validated using the *leave-one-out* approach. Where least squared regression functions were generated from random subsamples of the data leaving one example out each time.

Table 3.7 shows the attributes used for the regression, as well as some statistics (minimum, maximum, mean and standard deviation). The attributes were normalized before the prediction using the values of max_{depth} and max_{wn} shown in Table 3.1. The standard deviation shows a high stability for almost all the attributes. The attribute *log-hypo(lcs)* achieved the highest value of standard deviation.

Table 3.8 presents the predicted models when using the IIC and the IC_{hd} approaches. From there we can capture the importance of each attribute in the final prediction. How we can see the normalized logarithm of the number of hyponyms of the lowest common superset have the highest importance. In fact, the logarithm of the number of hyponyms of a concept, as a parameter, showed to be more important than the depth of the concept.

Analyzing the evaluation measures of the predicted models presented in Table 3.9 we can see the predicted model based on the IC_{hd} approach outperform the predicted model based on the IIC approach. The simple addition of the attribute *depth(lcs)* into the regression analysis for (IC_{hd}) make the predicted model to obtain better results than the IIC approach, as shown in Table 3.9. This result also comes to support our hypothesis.

Table B.14 in Appendix B contains the raw data submitted for the model prediction. The data have been presented in descending order depending on the value of the second last column (*log-hypo(lcs)*). If we look at the results in this table we can conclude that: when the number of hyponyms of the lowest common superset, (represented by is normalized value, *log-hypo(lcs)*), of the words under comparison decrease, i.e. tends to zero, the

Table 3.7: *Attributes used for the numeric prediction of information content metrics by using a linear regression and their descriptive analysis.*

Attributes	Description	Statistics			
		Minimum	Maximum	Mean	Std Dev
depth(c1)	Normalized value of the depth of concept c_1 in WordNet's taxonomy.	0	1	0.445	0.209
depth(c2)	Normalized value of the depth of concept c_2 in WordNet's taxonomy.	0	1	0.38	0.287
depth(lcs)	Normalized value of the depth of the lowest common superset of concepts c_1 and c_2 in WordNet's taxonomy.	0	1	0.313	0.22
log-hypo(c1)	Normalized value of the logarithm of the number of hyponyms of concept c_1 in WordNet's taxonomy.	0	1	0.259	0.246
log-hypo(c2)	Normalized value of the logarithm of the number of hyponyms of concept c_2 in WordNet's taxonomy.	0	1	0.228	0.236
log-hypo(lcs)	Normalized value of the logarithm of the number of hyponyms of the lowest common superset of concepts c_1 and c_2 in WordNet's taxonomy.	0	1	0.63	0.335
Human judgments	Human judgments of P&S experiment	0.393	3.43	1.54	1.003

Table 3.8: *Identifying the features which characterize the data. Predicted linear regression models.*

	Regression Model
Based on IIC approach	`Human judments` = $1.2522 * \texttt{log-hypo(c1)} + 0.2207 * \texttt{log-hypo(c2)} - 2.8571 * \texttt{log-hypo(lcs)} + 3.0138$
Based on IC_{hd} approach (v1)	`Human judments` = $-1.6073 * \texttt{depth(lcs)} + 1.1447 * \texttt{log-hypo(c1)} + 0.278 * \texttt{log-hypo(c2)} - 3.3936 * \texttt{log-hypo(lcs)} + 3.7677$
Based on IC_{hd} approach (v2)	`Human judments` = $-0.1843 * \texttt{depth(c1)} + 0.1885 * \texttt{depth(c2)} - 0.6002 * \texttt{depth(lcs)} + 0.6692 * \texttt{log-hypo(c1)} + 0.57 * \texttt{log-hypo(c2)} - 3.2891 * \texttt{log-hypo(lcs)} + 3.4969$

Table 3.9: *Evaluation measures of the regression models.*

Evaluation measures	IIC Approach	IC_{hd} Approach (v1)	IC_{hd} Approach (v2)
Correlation coefficient	0.8692	0.8762	0.8828
Mean absolute error	0.383	0.3459	0.3617
Root mean squared error	0.5032	0.4804	0.4723
Relative absolute error	42.407 %	38.3056 %	40.0539 %
Root relative squared error	49.7984%	47.537%	46.741 %
Total number of instances	65	65	65

words are very likely to be similar[3]. However, this behavior is only true when the value of the above mentioned attribute is very close to it's maximum value. Otherwise, the depth of the concepts and the depth of their *lcs* are important as we show in Table 3.10.

A detailed ranking analysis of the similarity results

During this experiment we compare the ranking generated by a similarity measure (Equation 4.17) when the IIC and the IC_{hd} approaches are used and try to make sense

[3]In our data just an exception was found, *hill* vs. *mound*, but presumably due to human disagreement since the words, in fact, are synonymous.

Table 3.10: *Snippet of the disagreement in the ranking comparison between the human judgment and the IIC and IC_{hd} approaches.*

			Distance from human ranking	
No.	word1	word2	using IIC	using IC_{hd}
1	cemetery#n#1	graveyard#n#1	3	4
2	automobile#n#1	car#n#1	**13**	**10**
3	coast#n#1	shore#n#1	4	3
4	boy#n#1	lad#n#2	10	11
5	journey#v#1	voyage#v#1	6	5
6	autograph#n#2	signature#n#1	1	3
7	cord#n#1	string#n#1	3	2
8	furnace#n#1	stove#n#1	12	13
9	brother#n#5	monk#n#1	12	11
10	sage#n#1	wizard#n#1	16	15
11	food#n#3	fruit#n#3	*26*	*33*
12	bird#n#1	cock#n#4	2	1
13	crane#n#5	rooster#n#1	3	2
14	cemetery#n#1	mound#n#1	28	29
15	car#n#1	journey#n#1	17	16
16	glass#n#2	jewel#n#1	1	2
17	furnace#n#1	implement#n#1	7	8
18	hill#n#1	woodland#n#1	**15**	**9**
19	food#n#1	rooster#n#1	1	3
20	shore#n#1	voyage#n#1	**23**	**20**
21	forest#n#2	graveyard#n#1	14	12
22	shore#n#1	woodland#n#1	**9**	**2**
23	lad#n#1	wizard#n#1	2	3
24	coast#n#1	forest#n#2	**8**	**4**
25	asylum#n#1	monk#n#1	8	6
26	cemetery#n#1	woodland#n#1	8	6
27	asylum#n#1	cemetery#n#1	8	7
28	boy#n#1	rooster#n#1	6	8
29	grin#n#1	lad#n#1	13	11
30	cushion#n#1	jewel#n#1	13	15
31	graveyard#n#1	madhouse#n#1	4	5
32	grin#n#1	implement#n#1	8	10
33	glass#n#5	magician#n#1	**12**	**7**
34	mound#n#1	stove#n#2	**15**	**3**
35	cord#n#2	smile#n#1	2	4
36	noon#n#1	string#n#4	1	3
37	fruit#n#2	furnace#n#1	**23**	**20**
38	asylum#n#1	fruit#n#2	**19**	**14**

out of it.

Table 3.10 is an snippet of Table B.15 in Appendix B where the ranked position of the similarity of each word pairs is compared with the ranking from human judgment for the two approach IIC and IC_{hd}. In this table we just show the pairs where the ranking from the two approach disagree.

Table 3.11 shows some statistics about the disagreements (or misplaced positions) extracted from Table B.15. The IC_{hd} approach have ranked the pair's similarity closer to the human ranking with higher frequency than the IIC approach (22 vs. 16). Form those 22 times the IC_{hd} approach ranked better than the IIC approach in 9 occasions (shown in bold in Table 3.10) the improvement was for more than 2 positions in the ranking versus just 1 occasion for the IIC approach (shown in italic in Table 3.10). The average and the standard deviation of the disagreements generated when using the IC_{hd} approach with respect to the human judgments are both smaller than the generated by the IIC approach, as also shown in Table 3.11. Up to this point we can say that when

Table 3.11: *Statistics from the disagreement (or misplaced positions) in the ranking comparison between the human judgment and the IIC and IC_{hd} approaches.*

Ranking comparison in R&G dataset	Disagreements	
	IC_{hd}	IIC
Pairs better ranked (w.r.t.)	22	16
Disagreements with difference (w.r.t) the other approach larger than 2	9	1
Disagreements Average	8.38	8.94
Disagreements Std Dev	7.33	7.36

the depth is also considered in the model (IC_{hd} approach) the ranking is closer to the humans' judgments. However a significance test for the means comparison of the two raking dataset showed that they are not significantly different, so we have not enough data to backup this statement.

3.5 General Discussion

The highest value of correlation when compared with the human judgment was obtained while using our proposed IC_{hd} in Lin's similarity (Sim_{lin}). Analyzing the obtained data we also realized of the competitiveness of our new information content metric compared to the intrinsic information content (IIC). Furthermore, it does outperform the traditional corpus based IC metric for all the similarity measures tested.

However, we still have to find out a better function for describing the relationship between the depth, the number of hyponyms and the information content value of a concept. Due to the "odd" behavior of the function *hypo* in the domain of English words with a wide range of possible values ($hypo(c) \in [0, 82115]$) it is difficult to properly model the relation between *IC*, *hypo* and *depth*. So, we have to model how does behave an information content metric given two concepts (c_1 and c_2) under the following conditions:

$$hypo(c_1) > hypo(c_2)$$
$$depth(c_1) < depth(c_2)$$

We tried to statistically approximate the metric expression through a linear regression but as we expected the relation between *hypo* and *depth* functions with the IC value it was not linear. So we followed a similar approach and approximated the metric with the *depth* and the logarithm of *hypo*. In that case our approach (IC_{hd}) also proved to be a better choice.

Another interesting point with our approach is related to the domain of the taxonomy. For some domains their associated knowledge's structure is uneven. When new knowledge is inserted, rather than growing up as a new level, they grow up but to the sides as a new sibling instead of a child. From this perspective new features related to the *width* of the taxonomy and the *number of siblings* a concept have, should be considered for computing the information content of a concept. Then it would be an open research in our approach to also include the width and the number of siblings in the formulation.

Zhou et. al. followed an approach similar to ours however there are some weaknesses in it [106]. Their expression depends on a parameter for the IC computation which have

to be empirically determined. However, the most serious of all is the expression they obtained models the behavior of the number of hyponyms and the depth of a concept as if they were similar features or having the same importance. The analysis of both features clearly shows this is not the case.

3.6 Chapter Conclusion

In this chapter we have shown a new approach for computing the information content values in a corpus independent way by using taxonomic properties. We proposed a new information content metric. This metric, when used in information content based semantic similarity measures, allows them to achieve better results than the traditional corpus based information content metric (IC). Our approach also solves the weakness of the intrinsic information content metric (IIC) while keeping the competitiveness of the metric.

Chapter 4

The Menendez-Ichise model

An ability to assess similarity lies close to the core of cognition. Its understanding supports the comprehension of human success in tasks like problem solving, categorization, memory retrieval, inductive reasoning, etc, and this is the main reason it is a common research topic.

In this chapter, we describe the motivation of the research (Section 4.1) and different abstract models of similarity (Section 4.2). We introduce the idea of semantic differences and commonalities between words to the similarity computation process, and propose a general model for it (Section 4.3). Five new semantic similarity metrics are obtained after applying this scheme to traditional WordNet-based measures. We also combine the node based similarity measures with a corpus-independent way of computing the information content. In an experimental evaluation of our approach (Section 4.4) on two standard word pairs datasets, four of the measures outperformed their classical version, while the other performed as well as their unmodified counterparts. A general discussion about the experimental results (Section 4.5) is also covered.

4.1 Motivation

In many fields such as artificial intelligence, biomedicine, linguistics, cognitive science, and psychology the semantic similarity of words is a topic of research. The computation of semantic similarity is extensively used in a variety of applications, like words sense disambiguation [75], detection and correction of malapropisms [9], information retrieval [88, 33], automatic hypertext linking [27] and natural language processing. Several applications to the field of artificial intelligence are discussed in [83]. However, despite numerous practical applications today, its theoretical foundations lies elsewhere, in cognitive science and psychology where it has been the subject of many investigations and theories (e.g [97, 86, 99, 26, 35]).

Let take a current example of peer-to-peer networks into which semantic similarity has found its way [29]. Assuming a shared taxonomy among the peers to which they can annotate their content, similarities among peers can be inferred by computing similarities among their representative concepts in the shared taxonomy. In this way, the more two peers are similar, the more efficient it is to route messages toward them. Numerous similar applications are the reasons for the increasing interest in this subject, whose ultimate goal is to mimic human judgment regarding similarity of word pairs.

As we previously mentioned, semantic similarity of words is often represented by the similarity between the concepts associated with the words. Several methods have been developed to compute word similarity, some of them operating on the taxonomic dictionary WordNet [24] and exploiting its hierarchical structure. However the majority of them suffer from a serious limitation. They only focus on the semantic information shared by those concepts, i.e., on the common points in the concept definitions or in the semantic differences but they never combine both. The increasing need for better measures and the new study area of semantic differences between words has led us to this study in the hope of upgrading existing semantic similarities measures.

In this chapter, we combined traditional WordNet-based semantic similarity measures with the idea of the "similarity between entities being related to their commonalities as well as to their differences", in order to improve the performance of WordNet-based sim-

ilarity measures and to obtain better results for applications using semantic similarities. Next subsections will cover the second group of WordNet-based semantic similarities, edge-based measures and introduce some recent node-based approaches of computing semantic relatedness which can be also used as semantic similarities.

4.1.1 Edge-based Semantic Similarity Measures

An intuitive way to quickly compute the semantic similarity between two nodes of a hierarchy is to count the number of edges in the shortest path between these two nodes. The idea behind this is that the semantic distance of two concepts is correlated with the length of the shortest path to join these concepts. This measure was first defined by Rada [72]. However, it relies upon the assumption that each edge carries the same amount of information, which is not true in most ontologies [74].

Many other formulas have since extended Rada's measure by computing weights on edges by using additional information, such as the depth of each concept in the hierarchy and the *lowest common superset, or most specific subsumer (lcs)* [102]. For example, in Figure 4.1, (where solid lines represent *IS-A* links and dashed lines indicate that some intervening nodes were omitted to save space), the *lcs* between the concepts *nickel* and *dime* is the concept *coin*.

The measures which focus on structural semantic information (i.e., the depth of the lowest common superset ($lcs(c_1, c_2)$), the depth of the concept's nodes, and the shortest path between them) are called *edge-based similarity measures*.

Wu and Palmer

In a paper on translating English verbs into Mandarin Chinese, Wu & Palmer introduced a scaled metric for what they call *conceptual similarity* between a pair of concepts c_1 and c_2 in a hierarchy [102].

Their strategy was to project verbs from both languages onto a common conceptual structure. Based on this conceptual representation, a similarity measure is defined that allows a target lexical item to be put in correspondence with a source item that

most closely carries the same meaning. The conceptual structures on to which verbs are projected are hierarchies (using hypernymy/hyponymy links). They proposed that the similarity between a pair of concepts c_1 and c_2 can be formulated as:

$$Sim_{wup}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{4.1}$$

where function $depth(c)$ represent the depth (or the length of the path from the root node of the hierarchy) of concept c, and $lcs(c_1, c_2)$ is lowest common superset (also called most specific subsumer or most specific common abstraction) of concepts c_1 and c_2. For example, in Figure 2.8, the *lcs* between the concepts *car* and *bicycle* is the concept *wheeled vehicle*.

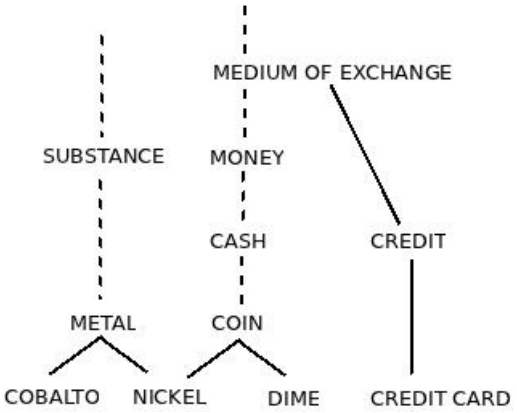

Figure 4.1: *Fragment of the WordNet taxonomy.*

Leacock and Chodorow

Leacock and Chodorow proposed a semantic similarity measure that typifies the edge-based approach [45]. In their measure, the similarity between two concepts is determined by first finding the length of the shortest path that connects them in the WordNet taxonomy. The length of the path that is found is scaled to a value between zero and one and similarity is then calculated as the negative logarithm of this value. The measure by Leacock and Chodorow may be expressed as follows:

$$Sim_{lch}(c_1, c_2) = -\log(\frac{length(c_1, c_2)}{2 * \lambda}) \tag{4.2}$$

where $length(c_1, c_2)$ denotes the length, counted in nodes, of the shortest path between the concepts c_1 and c_2; and λ denotes the maximum depth of the WordNet subsumption hierarchy.

The measure by Leacock and Chodorow can be illustrated with reference to the WordNet subgraph given in Figure 2.8. The shortest taxonomic path between *motorcycle* and *bicycle* is:

> *motorcycle IS-A motor vehicle IS-A self-propelled vehicle IS-A wheeled vehicle SUBSUMES bicycle*

It should be noted that the taxonomic path length differs from the network path length, as only hypernymy and hyponymy relations are considered. Assuming an arbitrary maximum depth of 10 in the WordNet taxonomy, the value of similarity between motorcycle and bicycle would be computed as:

$$\begin{aligned} Sim_{lch}(motorcycle, bicycle) &= -\log \frac{length(motorcycle, bicycle)}{2 \times 10} \\ &= -\log \frac{4}{20} \\ &= 0.6989 \end{aligned}$$

The Wu & Palmer and the Leacock & Chodorow similarity measures are based in a linear model, whereas Li et al.'s approach combines structural semantic information in a nonlinear model [47]. Li et al.'s model empirically defines a similarity measure that uses the shortest path length, depth, and local density in a taxonomy. They include two parameters which represent the contribution of the shortest path length and the depth of the *lcs* in the similarity computation process.

4.1.2 Other Semantic Similarity Measures

In Section 3.1.1 we introduced the node-based and the hybrid semantic similarity measures we used for this study. However, in this section we briefly present and discuss about some recently developed similarity measures which exploit not only WordNet but

the Web, the Linked Open Data (LOD)[1], and other taxonomies extracted from Wikipedia[2] and biomedical ontologies.

The FaITH (Feature and Information THeoretic) Similarity Measure

Pirró and Euzenat extended a previous work [67] to create a similarity measure which could also be used as a semantic relatedness and where different the *part-of-speech* described in WordNet could be considered [69]. They developed a featured and information theoretic based measure (Sim_{FaITH}) which use a ratio approach of the Tversky Abstract Model that we will introduce in the next section:

$$Sim_{FaITH}(c_1, c_2) = \frac{eIC(lcs(c_1, c_2))}{eIC(c_1) + eIC(c_2) - eIC(lcs(c_1, c_2))} \tag{4.3}$$

where eIC refers to Equation 3.16 introduced in the previous chapter, c_1 and c_2 represent the concepts, and $lcs(c_1, c_2)$ stands for the lowest common subsumer of concepts c_1 and c_2.

The W&IC Semantic Similarity Measure

Recently, Li et. al. proposed a variation of the Lin's similarity (Sim_{lin}) resulting in a very similar expression to the one obtained by Pirró and Euzenat [69] but considering a new parameter, the depth of the concept in the taxonomy [46]:

$$Sim_{W\&IC}(c_1, c_2) = \frac{IC(lcs(c_1, c_2))}{\lambda * IC(c_1) + (1 - \lambda) * IC(c_2)} \tag{4.4}$$

[1] http://linkeddata.org/
[2] http://www.wikipedia.org/

where λ is defined as follows:

$$\lambda = \begin{cases} \frac{depth(c_1)}{depth(c_1)+depth(c_2)} & if\ depth(c_1) \leq depth(c_2) \\ \\ 1 - \frac{depth(c_1)}{depth(c_1)+depth(c_2)} & if\ depth(c_1) < depth(c_2) \end{cases} \tag{4.5}$$

Web-based Semantic Similarity Measures

Sánchez et. al. proposed a web-based approach to compute semantic similarity which is applied to traditional WordNet information content based semantic similarity measures like Sim_{res}, Sim_{lin} and Sim_{jcn} [80]. For those measures a new way of computed the information content (see Equation 3.18) is used and later substituted in their original expressions.

However the roots of the Sánchez work can be tracked to the work of Bollegala et. al. where four other web-based semantic similarities were defined grounded on page count-based for popular co-occurrence measures; Jaccard, Overlap (Simpson), Dice and Pointwise Mutual Information (PMI) coefficients [8]:

$$Sim_{WebJaccard}(c_1, c_2) = \begin{cases} 0 & if\ H(c_1 \cap c_2) \leq t \\ \frac{H(c_1 \cap c_2)}{H(c_1)+H(c_2)-H(c_1 \cap c_2)} & otherwise \end{cases} \tag{4.6}$$

Therein, $H(c_1)$ represent the page counts for the query c_1 in a search engine, $c_1 \cap c_2$ denotes the conjunction query $c_1\ AND\ c_2$, and t is a threshold[3] for controlling non-related co-occurrences.

$$Sim_{WebOverlap}(c_1, c_2) = \begin{cases} 0 & if\ H(c_1 \cap c_2) \leq t \\ \frac{H(c_1 \cap c_2)}{\min(H(c_1), H(c_2))} & otherwise \end{cases} \tag{4.7}$$

$$Sim_{WebDice}(c_1, c_2) = \begin{cases} 0 & if\ H(c_1 \cap c_2) \leq t \\ \frac{2 * H(c_1 \cap c_2)}{H(c_1)+H(c_2)} & otherwise \end{cases} \tag{4.8}$$

[3] In their experiments they set $t = 5$

$$Sim_{WebPMI}(c_1, c_2) = \begin{cases} 0 & if\ H(c_1 \cap c_2) \leq t \\ \log_2(\frac{\frac{H(c_1 \cap c_2)}{N})}{\frac{H(c_1)}{N} * \frac{H(c_2)}{N}}) & otherwise \end{cases} \tag{4.9}$$

where N is the total number of documents indexed by the search engine.

Ontolgoy-based Semantic Similarity Measures

Sánchez et. al. also proposed an ontology-based to compute semantic similarity which is applied to traditional WordNet information content based semantic similarity measures like Sim_{res}, Sim_{lin} and Sim_{jcn} [78]. For those measures a new way of computed the information content (see Equation 3.17) is used and later substituted in their original expressions.

Batet et. al. proposed another ontology-based semantic similarity which was mainly applied into the biomedical domain [3, 4, 79]. The crafted similarity seems to be a modification of FaITH's similarity where a set-based approach is used, the numerator and the denominator have been exchanged and a logarithm is applied to the resulting fraction:

$$Sim_{batet}(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \tag{4.10}$$

where $T(c_i)$ stands for the union of the ancestors of the concept c_i and c_i itself. Although baptized as an ontology-based similarity the Batet's similarity can be also considered a set-based similarity.

Cross and Yu proposed an ontological similarity based on fuzzy set theory, information content and Tversky similarity [15]. The so-called Jaccard ontological similarity measure was defined as follows:

$$Sim_{JanAnc}(c_1, c_2) = \frac{\sum_{c \in F_{anc+(c_1)} \cap F_{anc+(c_2)}} IC(c)}{\sum_{c \in F_{anc+(c_1)} \cup F_{anc+(c_2)}} IC(c)} \tag{4.11}$$

where $F_{anc+}(c1)$ represents the fuzzy set of ancestors of concept c_1, the set intersection

uses a t-norm, typically minimum, and the set union uses a t-co-norm, typically maximum.

Cross and Yu said the *min* and *max* fuzzy set operators do not need to be explicitly used because the membership degrees of the ancestors in the fuzzy sets representing both concepts c_1 and c_2 are simply the ancestor's IC value. The IC value of the concept in each fuzzy set is the same since it is a function of its number of descendants. However, these IC values could be normalized so that the membership degree of an ancestor in each concept's fuzzy set could differ. For this approach, the *min* and *max* operators would then be needed since the IC membership degrees then differ. The ontological similarity could also be modified by describing a concept using a different set; for example, instead of the ancestor set to describe the concept, the descendant set could be used to describe each concept.

Another recently developed ontological approach was proposed by Saruladha et. al. and it is applied into the biomedical domain [81]. Their approach assesses semantic similarity among concepts from different and independent ontologies without constructing a priori a shared ontology. It is also based on Tversky similarity model and is mapped to information theoretic domain. They explored the possibility of adapting the existing single ontology information content based approaches and propose methods for assessing semantic similarity among concepts from different multiple ontologies. The proposed approaches have been experimented with two biomedical ontologies: SNOMED-CT (Systematized nomenclature of medical clinical terms) and Mesh (Medical subject headings) and the results were reported.

Saruladha's approach, based on the same grounds as Sánchez's [78] and Pirró's [69] approaches, obtained similar modified versions of traditional WordNet based semantic similarity but adding and information content variant which allowed the computation even when the concepts belong to different ontologies.

Although presented as a semantic relatedness Zhang et. al. also work out a model which can deal with several knowledge sources like WordNet and Wikipedia [104]. Their model probed to obtain good results in both the general and the biomedical domain, even though none domain-specific resource were used.

To finish with the ontology-based similarity measures we would like to mention the work of Dong et. al. [20]. The authors proposed a context-aware semantic similarity model for ontology environments. They stand when applying the semantic similarity model within a semantic-rich ontology environment, two issues are observed: (1) most of the models ignore the context of ontology concepts and (2) most of the models ignore the context of relations. In their paper they presented a solution for the two issues, including an ontology conversion process and a context-aware semantic similarity model, by considering the factors of both the context of concepts and relations, and the ontology structure. This approach implied an overhead delay for creating what they called the *lightweight ontology space* which can be done a priori (off-line) and it just need to be done once.

A Linked Open Data Based Approach

A very recent approach try to leverage the Linked Open Data (LOD) for computing semantic relatedness between named entities [105] . They said the existing knowledge based approaches have the entity coverage issue and the statistical based approaches have unreliable result to low frequent entities. LOD consists of lots of data sources from different domains and provides rich a priori knowledge about the entities in the world. By exploiting the semantic associations in LOD, they proposed a novel algorithm, called LODDO, to measure the semantic relatedness between named entities. This approach, although developed as a relatedness measure for named entities, can also be applied to compute semantic similarity between words.

4.2 Abstract Models of Similarity

Close to the core of cognition, similarity plays an indispensable foundational role in cognitive theories where several studies have been done. Four major psychological models of similarity are: geometric [97], featural [99], alignment-based [26] and transformational [35].

Geometric models have been among the most influential approaches to analyzing similarity. Geometric models standardly assume minimality [$D(A, B) \geq D(A, A) = 0$], symmetry [$D(A, B) = D(B, A)$], and the triangle inequality [$D(A, B) + D(B, C) \geq D(A, C)$]. Tversky criticized geometric models on the grounds that violations of all three assumptions are empirically observed [99].

When comparing things that are richly structured rather than just being a collection of coordinates or features often it is most efficient to represent things hierarchically (parts containing parts) and/or propositionally (relational predicates taking arguments). In such cases, comparing things involves not simply matching features, but determining which elements correspond to or align with one another.

In alignment-based models, matching features influence similarity more if they belong to parts that are placed in correspondence, and parts tend to be placed in correspondence if they have many features in common and if they are consistent with other emerging correspondences.

A fourth approach to modeling similarity is based on transformational distance. The similarity of two entities is assumed to be inversely proportional to the number of operations required to transform one entity so as to be identical to the other.

The key to calculating semantic similarity lies in resembling human thinking behavior. Semantic similarity of concepts is determined by processing first-hand information sources in the human brain. During this thesis we have introduced to several WordNet based semantic similarity measures. However, Table 4.1 shows a compilation of the different similarity measures we will work with in this chapter, as well as their main features.

For a better understanding of the foundations of the model presented in this paper we also introduce Tversky's abstract featural model of similarity [99].

4.2.1 Tversky Abstract Model of Similarity

In 1977, Tversky presented a model named the *Contrast Model* which takes into account features that are common to two concepts and features specific to each. That is, the similarity of concept c_1 to concept c_2 is a function of the features common to c_1 and

Table 4.1: *Compilation of the different semantic similarity measures and their main features.*

Type	Similarity	Description
Edge-based	Rada [72] Sim_{length}	Rely on the length of the shortest path joining two concepts.
	Wu & Palmer [102] Sim_{wup}	Rely on the depth of the lowest common superset between two concepts.
	Leacock & Chodorow [45] Sim_{lch}	Rely on the length of the shortest path between two synsets.
Node-based	Lin [48] Sim_{lin}	Defined by the ratio between the amount of information needed to state the commonality of the concepts and the information needed to fully describe what the concepts are.
	Resnik [74] Sim_{res}	Defined by the information content of the lowest common superset between two concepts.
	Pirró & Seco [70] $Sim_{p\&s}$	Based on Tversky's theory but from an information theoretic approach.
Hybrid	Jiang & Conrath [38] $Sim_{j\&c}$	A combined approach where the edge counting scheme is enhanced by the information content approach.

c_2, those in c_1 but not in c_2 and those in c_2 but not in c_1. Admitting a function $\psi(c)$ that yields the set of features relevant to c, he proposed the following similarity function:

$$Sim_{tvr}(c_1, c_2) = \alpha F(\psi(c_1) \cap \psi(c_2)) - \beta F(\psi(c_1)/\psi(c_2)) - \gamma F(\psi(c_2)/\psi(c_1)) \tag{4.12}$$

where F is some function that reflects the salience of a set of features, and α, β and γ are parameters provided for differences in each component. According to Tversky, similarity is not symmetric, that is, $Sim_{tvr}(c_1, c_2) \neq Sim_{tvr}(c_2, c_1)$, because humans tend to focus more on one object than on the other depending on the way the relationship direction is taken into consideration during the comparison.

For example, regarding the concept *dime* in Figure 4.1, which represent a fragment of the WordNet taxonomy adapted from [74], it is logical that one of it's most related concepts is *nickel*, but the same is not true in the opposite direction. The concept *nickel* is also like *cobalt, gold, metal, etc.*

4.3 Semantic Commonalities and Differences in Similarity Measures

Most of the WordNet-based semantic similarity measures just take into consideration semantic commonalities among concepts for computing their values. The strength of semantic differences has been diminished or not fully exploited while their combination have been rarely considered from a broader perspective. Having all these elements in mind and considering the current structure of WordNet taxonomy, we propose the Menendez-Ichise model [55].

In this section, we introduce our model and its application to traditional WordNet based similarity metrics. The modifications to those metrics are founded on Tversky's Contrast Model theory of similarity [99] which is classified as featural model of similarity.

Our model supports to be a specialization of Tversky's featured-based theory applied to traditional WordNet-based semantic similarity measures. Paraphrasing Tversky, we state that: "the similarity between two entities is related to their commonalities as well as to their differences," and our general model is described by the following expression:

$$Sim(c_1, c_2) = \alpha * Comm(c_1, c_2) - \beta * Diff(c_1, c_2) \tag{4.13}$$

where $Comm(c_1, c_2)$ stands for ***commonalities***, $Diff(c_1, c_2)$ for the ***differences***, and α and β are tuning factors $(0 \leq \alpha)$ and $(0 \leq \beta)$ that represent the importance of the commonalities and differences in the model. Because WordNet's structure is represented by an undirected graph we can't avoid assuming symmetry where there is none.

The use of semantic differences for computing semantic similarity and its combination with the semantic commonalities is a novel approach. Below we explain how we applied our model to WordNet-based semantic similarity measures.

4.3.1 Application of the Menendez-Ichise Model

The main features considered by WordNet-based similarity metrics are, the distance between nodes and the weight of the nodes. This in turn leads to two different approaches: *edge-based* and *node-based*, as mentioned above.

In the Menendez-Ichise model, regardless of the approach used, we consider the information from the *root*[4] to the $lcs(c_1, c_2)$ as the ***semantic commonalities*** of the concepts c_1 and c_2; and the rest of the information from the $lcs(c_1, c_2)$ to each of the concepts c_1 and c_2 as the ***semantic differences***. Hence, from the perspective of an edge-based approach, the differences are related to the shortest path between the two concepts while the commonalities are related to the depth of the *lcs*, in other words, to the path from the *root* to the *lcs*. In node-based approach, the differences are related to the information contained in the nodes representing the concepts but not contained in their *lcs*, because this last one its encapsulating the common information.

For example, regarding the concepts *nickel* and *dime* in Figure 4.1, the semantic commonalities are in their *lcs*, i.e, the taxonomy subgraph from the *root* to the $lcs(nickel, dime) = coin$. The semantic differences between both concepts is enclosed in the taxonomy subgraph from $lcs(nickel, dime)$ to both concepts but without considering any information from the *root* to the concept *coin*.

Modified Length Similarity Measure

Equation 4.14 is a combination of the traditional length and depth of the *lcs* metrics, because each of them deal with the differences and the commonalities respectively. We consider the first term of Sim'_{length} [5]as the semantic commonalities between the concepts, which is twice the distance from the *root* to their *lcs*. The second term as the semantic differences in this case the distance between the two concepts.

[4]The most abstract node in the taxonomy.

[5]From now on this notation Sim'_{abbr} will be used for representing the modified similarity expression corresponding to the authors or model specified by *abbr*.

$$Sim'_{length}(c_1, c_2) = \alpha * \left(1 - \frac{1}{2 * depth(lcs(c_1, c_2))}\right) - \beta * \left(1 - \frac{1}{length(c_1, c_2) + 1}\right) \tag{4.14}$$

$depth(c)$: depth of concept c in the taxonomy, where the depth of the most abstract node, the "root", is 1.

$length(c_1, c_2)$: number of edges from node c_1 to node c_2 in the taxonomy.

Modified Wu & Palmer and Leacock & Chodorow Similarity Measures

While Wu & Palmer measure rely on the depth of the lowest common superset between the concepts (*semantic commonalities*), the Leacock & Chodorow measure rely on the length of the shortest path between two synsets (*semantic differences*). Now for each of their modified expressions (Equation 4.15 and Equation 4.16) we have considered both the *semantic differences* and the *semantic commonalities*; which were not taken into consideration in their original formulation.

To follow the approach of their original expressions the commonalities and the differences have been normalized using a different approach for each case. While Wu & Palmer measure used a *normalization factor* (the addition of the concepts' depths in the taxonomy), Leacock & Chodorow metric, used the properties of the the logarithm function to soften its values after dividing by twice the taxonomy's depth.

$$Sim'_{wup}(c_1, c_2) = \alpha * \left(\frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}\right) - \beta * \left(\frac{length(c_1, c_2)}{depth(c_1) + depth(c_2)}\right) \tag{4.15}$$

$$Sim'_{lch}(c_1, c_2) = \alpha * \left(-\log \left(\frac{depth(lcs(c_1, c_2))}{2 * \lambda} \right) \right)$$

$$-\beta * \left(-\log \left(\frac{length(c_1, c_2)}{2 * \lambda} \right) \right) \tag{4.16}$$

λ : maximum depth of the taxonomy.

Modified Resnik Similarity Measure

The modified Resnik's similarity measure $Sim'_{res}(c_1, c_2)$ considers the *semantic commonalities* to be the information content of the $lcs(c_1, c_2)$ and the *semantic differences* to be the information content encompassed by concepts, minus the one already considered in the $lcs(c_1, c_2)$.

$$\begin{aligned} Sim'_{res}(c_1, c_2) = \alpha * IC(lcs(c_1, c_2)) \\ -\beta * (IC(c_1) + IC(c_2) - 2 * IC(lcs(c_1, c_2))) \end{aligned} \tag{4.17}$$

After the application of our model, the modified Jiang & Conrath similarity expression $Sim'_{j\&c}(c_1, c_2)$ is identical to the one obtained for Resnik's measure, Equation 4.17, and it is a generalization of the Pirró & Seco similarity measure, Equation 3.11.

$$Sim_{P\&S} \subset Sim'_{Res}(c_1, c_2) = Sim'_{j\&c}(c_1, c_2) \tag{4.18}$$

Modified Lin Similarity Measure

According to Lin "the similarity between c_1 and c_2 is measured by the ratio between the amount of information needed to state the commonality of c_1 and c_2 and the information needed to fully describe what c_1 and c_2 are" [48]. In Equation 4.19, we add the *semantic differences* as the information content in each concept minus the one already considered in the $lcs(c_1, c_2)$ divided by the information needed to fully describe the concepts. For Lin's expression the information needed to fully describe the concepts becomes a *normalization*

factor (see Table 4.2) whose effect we will discuss later.

$$Sim'_{lin}(c_1, c_2) = \alpha * \left(\frac{2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \right) - \beta * \left(\frac{IC(c_1) + IC(c_2) - 2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \right) \tag{4.19}$$

Table 4.2: *Normalization factor used with different metric approaches.*

Metric	Approach	Normalization Factor
Sim'_{wup}	edge-based	$depth(c_1) + depth(c_2)$
Sim'_{lin}	node-based	$IC(c_1) + IC(c_2)$

4.4 Experiments and Results

4.4.1 Experimental Settings

The purpose of the experiments is to prove the hypothesis that the use of semantics commonalities as well as semantics differences can improve the computation of similarity between concepts. We also want to test the effectiveness of the IC_{hd} corpora independent information content metric with node-based similarity measures. In the experiments we evaluate the new semantic similarity measures and establish a baseline for comparison of their results with those of their original versions.

Unfortunately, there is a distinct lack of standards for evaluating semantic similarities. This means the accuracy of a computational method for evaluating word similarity can only be established by comparing its results against human common sense. That is, a method that comes close to matching human judgments can be deemed accurate.

The procedure for evaluating the quality of the developed similarity measures its described in Figure 4.2. In general, after choosing the dataset of word pairs, we will compute the similarity between the words using different similarity measures. To deal with the polysemy property of words, the similarity for each possible combination of meaning for each word pair will be computed. We keep the pair of concepts whose

```
-----------------------------------------------

select dataset of word pairs;
for each similarity measure (Sim_i) {
  for each pair of concepts (c1_j,c2_j) in
      dataset {
    if (Sim_i is an edge-based similarity) {
      compute Sim_i(c1_j,c2_j);
    }
    else {
      // compute the similarity value using
      // different  information content
      // metrics (IC, IIC and IC_hd)

      for each information content metric
        compute Sim_i(c1_j,c2_j);
    }
  }

  //compute correlation factor
  Corr_Sim_i =
     Correlation_Factor(Sim_i,Human_Judgment);

  // compare Corr_Sim_i with the correlation
  // of the original similarity measure version
  compare(Corr_Sim_i, Corr_Orig_Version);
}
-----------------------------------------------
```

Figure 4.2: *Algorithm for evaluating the quality of the similarity measures*

similarity is maximal in the previous step. Each node-based similarity measures will be computed using three different information content metrics (IC, IIC, IC_{hd}). In all the measures the importance of the commonalities and the differences will be changed to assess the best ratio between them.

We followed the human judgments of Pirró and Seco experiment for the word pairs of both datasets, the Miller and Charles dataset (M&C) and the Rubenstein and Goodenough dataset (R&G) [70]. Again in M&C dataset we considered only 28 word pairs of the 30 used in the M&C experiment since a word missing in WordNet 3.0 made it impossible to compute ratings for the other two word pairs.

All the evaluations were performed using WordNet 3.0 [24] and the Brown Corpus[6]

[6]The Brown University Standard Corpus of Present-Day American English.

was used for the calculation of the corpus-dependent information content metric. For the computation we used Pedersen's WordNet::Similarity Perl module as the core. We also recreated the Pirró and Seco's experiment with the Java WordNet Similarity Library [70] (JWSL) using Pirró and Seco's intrinsic information content (IIC), but we did not obtain the same results they did. Probably due to the selection of the parameters. Table 3.1 (introduced in the previous chapter) shows the values we used for the parameters during the computation of the corpus independent information content metric (IIC and IC_{hd}) for the nouns and verbs WordNet's sub-taxonomies.

For the each metric we performed two experiments. The purpose of the first experiment is to check if the semantics differences have any effect(positive or negative) in the performance of the measures when they are also considered in the computation. In the second experiment we pursue to narrow the values for α and β which generate the higher performance of the semantic similarity measures. In other words to narrow the ratio of influence between the semantic differences and the semantic commonalities in the semantic similarity computation.

In the first experiment we variated the importance of the semantic differences' factor, β, in the range $[0, 1]$ while keeping constant the value of importance of the semantic commonalities ($\alpha = 1$). Then we calculate the Pearson correlation of the new metrics' results with respect to the human judgments values obtained in [70]. In the second experiment we do variate the importance of the semantic differences with values in the same range as in the first experiment but we also variate the importance of the semantic commonalities.

We assumed both factors cooperate for the final goal, so we set $\alpha = 1 - \beta$. In both experiments we used an step of 0.01 for the variation of β. We also experimented we values of α and β in the range $[-100, 100]$ but the best results were obtained when both factors were in the $[0, 1]$ range.

Table 4.3 shows the general details for each experiment. In both experiments, for the node-based measures we also check the effectiveness of the IC_{hd} information content.

Table 4.3: *The Menendez-Ichise Model. Experiments description.*

Experiment	α	β
Exp. 1	$\alpha = 1$	$0 \leq \beta \leq 1$
Exp. 2	$\alpha = 1 - \beta$	$0 \leq \beta \leq 1$

4.4.2 Experimental Results

Experiment 1

Table 4.4 compiles the results of the first experiment for the edge-based similarity measures using several values for the *differences' factors* for the M&C dataset. We did not include in the table all the values of β used to run the experiment, just a representative sample for values of it. This means sometimes we could have values of β for which a higher correlation value was obtained but they will be shown at the end in the experiment's summary. The second column, entitled "Original", represents the results of the original measure[7], i.e., the previous or baseline result. The correlation value for the unmodified functions and when $\beta = 0.0$ would be the same if the modified measure considers the commonalities as in the original metric.

Table 4.4: *Correlation coefficients obtained for edge-based measures in Exp. 1 using the 28 words' pairs of M&C dataset.*

		β				
	Original	0.0	0.1	0.3	0.6	1.0
Sim'_{length}	0.8401	0.6673	0.7958	0.8498	**0.8571**	0.8549
Sim'_{wup}	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**
Sim'_{lch}	**0.8293**	-0.7126	-0.7446	-0.7804	-0.8039	-0.8165

This is not the case for Sim'_{length} and Sim'_{lch} similarity measures and it is the reason for the difference in the correlation values between the original function and the modified version when $\beta = 0.0$. But going deeper in the details of the results we can say:

1. Sim'_{length} effectiveness improved when the semantic differences were considered (compared with path-length and depth of the *lcs* metrics). The ratio between the semantic differences and the semantic commonalities ($\frac{\beta}{\alpha} = 0.6$) was 0.6.

[7]The original metric has not been modified.

2. Sim'_{wup} effectiveness remains the same as its original version, showing no changes for any value of the semantic differences' importance, β. The normalization done for this measure generated an expression which, rather than similar to the "Contrast Model" of the Tversky's feature-based approach [99], get closer to the "Ratio Model" of Tversky's abstract approach.

3. Sim'_{lch} did not improve its correlation value when compared with its original expression. Although when increasing the values of β the Sim'_{lch} expression was achieving better results, with values of $0 \leq \beta \leq 1$ and $\alpha = 1$, the original expression behaved better.

Table 4.5 compiles the results of the first experiment for the edge-based similarity measures using a larger dataset of words' pairs, the R&G dataset. The general description for Table 4.4 is also valid in here but let go through the details:

Table 4.5: *Correlation coefficients obtained for edge-based measures in Exp. 1 using the 65 words' pairs of R&G dataset.*

		β				
	Original	0.0	0.1	0.3	0.6	1.0
Sim'_{length}	0.8373	0.4424	0.5974	0.7504	0.8200	**0.8420**
Sim'_{wup}	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**
Sim'_{lch}	**0.8631**	-0.6604	-0.7126	-0.7753	-0.8189	-0.8426

1. Sim'_{length} effectiveness improved when the semantic differences were considered. For this dataset (R&G) this measure improved with a higher value of importance for the semantic difference than for the M&C dataset. However, the correlation value was smaller in this occasion.

2. Despite the different dataset, the modified expression of Wu & Palmer (Sim'_{wup}) remains the same as its original version showing no changes for any value of β.

3. Sim'_{lch} did not improve its correlation value when compared with its original expression. Same behavior as observed for the M&C word pairs dataset.

Table 4.6 compiles the results of the Exp. 1 for the node-based similarity measures using several *differences' factors* and three different information content metrics: IC, IIC and IC_{hd} for the M&C words' pairs dataset. Again the column entitled "Original" represents the results of the original measure. The correlation value for the unmodified expression when $\beta = 0.0$ for $Sim'_{j\&c}$ measure is different because in $Sim'_{j\&c}$ the commonalities were not consider as in the original metric. $Sim_{p\&s}$ was not included because it is not affected when changing the values of β. In any case, we already probed Sim'_{res} is a more general expression than $Sim_{p\&s}$, Equation 4.18. After a deeper analysis of the results we can say:

Table 4.6: *Correlation coefficients obtained for node-based measures in Exp. 1 using the 28 words' pairs of M&C dataset and three different information content metrics.*

			β			
		Original	0.0	0.3	0.6	1.0
Sim'_{lin}	IC	**0.8587**	**0.8587**	**0.8587**	**0.8587**	**0.8587**
	IIC	**0.8797**	**0.8797**	**0.8797**	**0.8797**	**0.8797**
	IC_{hd}	**0.8821**	**0.8821**	**0.8821**	**0.8821**	**0.8821**
Sim'_{res}	IC	0.8308	0.8308	0.8555	0.8624	**0.8655**
	IIC	0.8421	0.8421	0.8740	0.8816	**0.8843**
	IC_{hd}	0.8361	0.8361	0.8743	0.8819	**0.8835**
$Sim'_{j\&c}$	IC	-0.8660	0.8308	0.8555	0.8624	**0.8655**
	IIC	-0.8805	0.8421	0.8740	0.8816	**0.8843**
	IC_{hd}	-0.8712	0.8361	0.8743	0.8819	**0.8835**

1. Sim'_{lin} achieved its highest value when combined with the IC_{hd} metric. But a similar behavior to the one observed for Sim'_{wup} was showed by Sim'_{lin} which no matter the value of the importance factors for the semantics differences, it remains the same as its original version as result of the normalization.

2. Sim'_{res} measure obtained higher values of correlation than the original expression when the semantic differences were considered, showing the best results for $\beta = 1.0$. The results of the measure when combined with IC_{hd} approach overcome the IC approach while remain as competitive as with the IIC approach.

3. $Sim'_{j\&c}$ also obtained higher values than its original expression, although only for the corpus independent information content metrics (IIC, IC_{hd}). The negative value of

the original function is due to the reason that the original expression is a distance and not a similarity. Since the expression for $Sim'_{j\&c}$ is equal to Sim'_{res} the rest of the conclusions are the same as above.

Table 4.7 compiles the results of the Exp. 1 for the node-based similarity measures using a larger dataset of word pairs (R&G):

Table 4.7: *Correlation coefficients obtained for node-based measures in Exp. 1 using the 65 words' pairs of R&G dataset and three different information content metrics.*

			β			
		Original	0.0	0.3	0.6	1.0
Sim'_{lin}	IC	**0.8812**	**0.8812**	**0.8812**	**0.8812**	**0.8812**
	IIC	**0.8992**	**0.8992**	**0.8992**	**0.8992**	**0.8992**
	IC_{hd}	**0.9007**	**0.9007**	**0.9007**	**0.9007**	**0.9007**
Sim'_{res}	IC	0.8677	0.8677	0.8792	**0.8802**	0.8792
	IIC	0.8773	0.8773	0.8928	**0.8949**	0.8944
	IC_{hd}	0.8679	0.8679	0.8903	**0.8927**	0.8915
$Sim'_{j\&c}$	IC	-0.8689	0.8677	0.8792	**0.8802**	0.8792
	IIC	-0.8848	0.8773	0.8928	**0.8949**	0.8944
	IC_{hd}	-0.8747	0.8679	0.8903	**0.8927**	0.8915

1. For this larger dataset Sim'_{lin} improved its effectiveness when combined with IC_{hd} approach but it was not affected by changing the importance of the semantic differences.

2. Sim'_{res} measure improved its effectiveness compared with its original expression when our model is applied. Again the combination with IC_{hd} approach overcome the IC approach while remain as competitive as with the IIC approach. For this larger dataset the ratio between the semantic differences and the semantic commonalities showed certain stability in the vicinity of 0.6. However, the maximum correlation values were obtained for the following values of β: 0.65 for IC, 0.75 for IIC and 0.63 for IC_{hd}.

3. $Sim'_{j\&c}$ achieved better results than its original expression. Same conclusions obtained for Sim'_{res} also applied to it.

From what we have seen so far, Exp. 1 supports our hypothesis, semantic differences are important for computing the semantic similarity between words.

Experiment 2

In Exp. 2 we want to investigate which is the ratio of importance between semantic commonalities and semantic differences. Then the values of β and α variate during the experiment.

Table 4.8 compiles the results of the second experiment for the edge-based similarity measures using several values for the *differences' factors* for the M&C dataset. We did not include in the table all the values of β used to run the experiment, just the most representative. The general description for Table 4.4 is also valid in here but let go through the details:

Table 4.8: *Correlation coefficients obtained for edge-based measures in Exp. 2 using the 28 words' pairs of M&C dataset.*

		β				
	Original	0.0	0.1	0.3	0.6	1.0
Sim'_{length}	0.8401	0.6673	0.8027	**0.8557**	0.8518	0.8401
Sim'_{wup}	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**
Sim'_{lch}	**0.8293**	-0.7126	-0.7474	-0.7931	-0.8229	**-0.8293**

1. Sim'_{length} effectiveness improved when the semantic differences were considered (compared with path-length and depth of the *lcs* metrics). The maximum value of correlation (0.8571) was obtained for $\beta = 0.37$. However this result shows for this similarity the commonalities have higher importance than the differences.

2. Sim'_{wup} behaved as in Table 4.4. Later on we will analyze in detail the reason of this behavior.

3. Sim'_{lch} slightly improved its correlation value when compared with its original expression since their absolute value is closer to 1. The highest value of correlation (-0.8296) was obtained for $\beta = 0.93$. The correlation values of the modified version

are negative confirming that our approach for considering the semantic commonalities and semantic differences is opposed to the approach used in the original measure, so it represents an inverse correlation. In Sim'_{lch} the semantic differences were considered a negative element in the expression while in the original formulation it was considered a positive element. Highly related or similar concepts obtained low values while unrelated concepts obtained high values. From this perspective the modified version is behaving like a distance rather than a similarity, therefore we could do the comparison using the absolute values of the correlation.

For this similarity measure the ratio between the semantic differences and the semantic commonalities ($\frac{\beta}{\alpha} = \frac{0.93}{0.07} = 13.29$) shows the semantic differences are (a lot) more important than the commonalities for the similarity computation process.

Table 4.9 compiles the results of the second experiment for the edge-based similarity measures using a larger dataset of words' pairs, the R&G dataset. The general description for Table 4.4 is also valid in here but let go through the details:

Table 4.9: *Correlation coefficients obtained for edge-based measures in Exp. 2 using the 65 words' pairs of R&G dataset.*

		β				
	Original	0.0	0.1	0.3	0.6	1.0
Sim'_{length}	0.8373	0.4424	0.6106	0.7921	**0.8476**	0.8373
Sim'_{wup}	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**
Sim'_{lch}	**0.8631**	-0.6604	-0.7174	-0.7987	-0.8542	**-0.8631**

1. Sim'_{length} effectiveness improved when the semantic differences were considered. The best correlation value (0.8481) was obtained for $\beta = 0.66$. The ratio between the semantic differences and the semantic commonalities ($\frac{\beta}{\alpha} = 1.94$) which suggests the importance of the semantic differences is higher than the importance of the commonalities when we have a larger dataset like R&G. This result is opposed to what we obtained from the same experiment with the smaller dataset M&C.

2. Despite the different dataset, the modified expression Sim'_{wup} remained the same as its original version showing no changes for any value of β.

3. Sim'_{lch} slightly improved its correlation value compared to the original version, since their absolute value is closer to 1. The best correlation value (0.8647) was obtained for $\beta = 0.87$. The semantic differences seem to be ($\frac{\beta}{\alpha} = 6.69$) times more important than the commonalities for the similarity computation. Although this value is smaller than the one obtained for the M&C dataset, the important point is that for Sim'_{lch} no matter which dataset was used, the semantic differences are more important than the semantic commonalities.

Table 4.10 compiles the results of the Exp. 2 for the node-based similarity measures using several *differences' factors* and three different information content metrics: IC, IIC and IC_{hd} for the M&C words' pairs dataset. Description for Table 4.6 also apply in here:

Table 4.10: *Correlation coefficients obtained for node-based measures in Exp. 2 using the 28 words' pairs of M&C dataset and three different information content metrics.*

			β			
		Original	0.0	0.3	0.6	1.0
Sim'_{lin}	IC	**0.8587**	**0.8587**	**0.8587**	**0.8587**	**0.8587**
	IIC	**0.8797**	**0.8797**	**0.8797**	**0.8797**	**0.8797**
	IC_{hd}	**0.8821**	**0.8821**	**0.8821**	**0.8821**	**0.8821**
Sim'_{res}	IC	0.8308	0.8308	0.8594	**0.8667**	0.8660
	IIC	0.8421	0.8421	0.8784	**0.8849**	0.8805
	IC_{hd}	0.8361	0.8361	0.8787	**0.8820**	0.8699
$Sim'_{j\&c}$	IC	-0.8660	0.8308	0.8594	**0.8667**	0.8660
	IIC	-0.8805	0.8421	0.8784	**0.8849**	0.8805
	IC_{hd}	-0.8712	0.8361	0.8787	**0.8820**	0.8699

1. Sim'_{lin} behaved as in Table 4.6. Same conclusions can be applied in here.

2. Sim'_{res} measure obtained higher values of correlation than the original expression when the semantic differences were considered. For this dataset the ratio between the semantic differences and the semantic commonalities showed certain stability in the vicinity of 0.6. However, the maximum correlation values for the information content metrics IC (0.8672), IIC(0.8849) and IC_{hd}(0.8835) were obtained for the following values of β: 0.76 for IC, 0.61 for IIC and 0.49 for IC_{hd}. Except when the

IC_{hd} approach was used for the other two (IC and IIC) the semantic differences have higher relevance than the commonalities.

3. $Sim'_{j\&c}$ behaved the same as Sim'_{res}. Same conclusions can be applied in here.

Table 4.11 compiles the results of the Exp. 2 for the node-based similarity measures using a larger dataset of word pairs (R&G):

Table 4.11: *Correlation coefficients obtained for node-based measures in Exp. 2 using the 65 words' pairs of R&G dataset and three different information content metrics.*

			β			
		Original	0.0	0.3	0.6	1.0
Sim'_{lin}	IC	**0.8812**	**0.8812**	**0.8812**	**0.8812**	**0.8812**
	IIC	**0.8992**	**0.8992**	**0.8992**	**0.8992**	**0.8992**
	IC_{hd}	**0.9007**	**0.9007**	**0.9007**	**0.9007**	**0.9007**
Sim'_{res}	IC	0.8677	0.8677	**0.8805**	0.8785	0.8699
	IIC	0.8773	0.8773	**0.8942**	0.8931	0.8848
	IC_{hd}	0.8679	0.8679	**0.8921**	0.8888	0.8740
$Sim'_{j\&c}$	IC	-0.8689	0.8677	**0.8805**	0.8785	0.8699
	IIC	-0.8848	0.8773	**0.8942**	0.8931	0.8848
	IC_{hd}	-0.8747	0.8679	**0.8921**	0.8888	0.8740

1. Sim'_{lin} obtained similar results as in Exp. 1, see Table 4.7. The behavior of this similarity will be analyzed in details in the next section.

2. Sim'_{res} measure obtained same maximum results as in in Experiment 1. However, for this dataset the ratio between the semantic differences and the semantic commonalities showed certain stability when β was in the vicinity of 0.3. The maximum correlation values for the information content metrics IC (0.8808), IIC(0.8949) and IC_{hd}(0.8927) were obtained for the following values of β: 0.38 for IC, 0.41 for IIC and 0.38 for IC_{hd}. For this larger dataset the values of β showed a greater importance for the semantic commonalities than for the semantic differences.

3. $Sim'_{j\&c}$ measure behaved the same as Sim'_{res}.

Summary of Experiments

In this section we compile the results of the experiments for both M&C and R&G datasets.

Table 4.12 shows the maximum values of correlation obtained for Sim'_{length}, Sim'_{wup} and Sim'_{lch} measures from which we arrived to the following conclusions:

Table 4.12: *Maximum values of correlation obtained for Sim'_{length},Sim'_{wup} and Sim'_{lch} measures using M&C and R&G datasets.*

	Pairs	Correlation		Parameter
	Dataset	Original	Max.	β
Sim'_{length}	M&C	0.8401	**0.8571**	0.37
	R&G	0.8373	**0.8481**	0.66
Sim'_{wup}	M&C	**0.7726**	**0.7726**	-
	R&G	**0.7795**	**0.7795**	-
Sim'_{lch}	M&C	0.8293	**-0.8296**	0.93
	R&G	0.8631	**-0.8647**	0.87

1. For both datasets Sim'_{length} obtained higher values when our model is applied. The correlation with respect to the human judgment was always higher for the M&C dataset. The importance of the semantic differences with respect to the semantic commonalities changed for both datasets. For the R&G dataset the semantic differences had higher importance than the commonalities, the opposed for the M&C dataset.

2. After the application of our model the Sim'_{lch} slightly improved its correlation values for both datasets. For both datasets the ratio between the semantic differences and the semantic commonalities ($\frac{\beta}{\alpha}$) showed higher importance for the semantic differences. This result its due to the original construction of the function where the differences had the leading vote.

3. Sim'_{wup} measure is not affected either positively or negatively from our model. Now the question is: why?

   When we applied our model to obtain Equation 4.15 we assumed the denominator of the original Sim_{wup} equation was seeking a normalization behavior for the final

result. So, let reduce the equation we obtained for Sim'_{wup}:

$$Sim'_{wup}(c_1, c_2) = \alpha * \left(\frac{2*depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}\right) - \beta * \left(\frac{length(c_1,c_2)}{depth(c_1)+depth(c_2)}\right)$$

`substituing:` $length(c_1, c_2) = depth(c_1) + depth(c_2) - 2 * depth(lcs(c_1, c_2))$

$$Sim'_{wup}(c_1, c_2) = \alpha * \left(\frac{2*depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}\right) - \beta * \left(\frac{depth(c_1)+depth(c_2)-2*depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}\right)$$

`reducing`

$$Sim'_{wup}(c_1, c_2) = 2 * (\alpha + \beta) * \left(\frac{depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}\right) - \beta * \left(\frac{depth(c_1)+depth(c_2)}{depth(c_1)+depth(c_2)}\right)$$

`substituing:` $\alpha + \beta = 1$

`reducing`

$$Sim'_{wup}(c_1, c_2) = \left(\frac{2*depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}\right) - \beta$$

In fact, this expression is almost (with the exception of the β parameter) the exactly same expression as the original Sim_{wup}, Equation 4.1. There is the reason why the correlation values were not affected when the value of β was changed. As we mentioned before, it seems the Sim_{wup} have been expressed using the ratio approach of the Tversky's abstract model of similarity. This exactly same analysis applied for the modified version of Lin's similarity.

In Table 4.13 we show Sim'_{lin} and $Sim_{p\&s}$ measures. The $Sim_{p\&s}$ measure was included as a reference. The Sim'_{lin} measure as Sim'_{wup} is not affected by the application of our model. However, its results considering various information content metrics were included. The highest value of correlation among all the studied similarities was achieved by Sim_{lin} when combined with our IC_{hd} approach.

Table 4.14 compiles the results for Sim'_{res} after the application of our model. The results for $Sim'_{j\&c}$ were also included for reference. From this table we can arrive to the

Table 4.13: *Maximum values of correlation obtained for Sim'_{lin} and $Sim_{p\&s}$ using M&C and R&G datasets.*

	Pairs DS	IC metric		
		IC	IIC	IC_{hd}
Sim'_{lin}	M&C	0.8587	0.8797	**0.8821**
	R&G	0.8812	0.8992	**0.9007**
$Sim_{p\&s}$	M&C	0.8655	**0.8843**	0.8835
	R&G	0.8793	**0.8944**	0.8915

following conclusions:

Table 4.14: *Maximum values of correlation obtained for Sim'_{res} and $Sim'_{j\&c}$ measures using different IC metrics in M&C and R&G datasets.*

	Pairs Dataset	IC Metric	Correlation		Parameter
			Original	Max	β
Sim'_{res}	M&C	IC	0.8308	**0.8672**	0.76
		IIC	0.8421	**0.8849**	0.61
		IC_{hd}	0.8361	**0.8835**	0.49
	R&G	IC	0.8677	**0.8808**	0.38
		IIC	0.8773	**0.8949**	0.41
		IC_{hd}	0.8679	**0.8928**	0.38
$Sim'_{j\&c}$	M&C	IC	-0.8660	**0.8672**	0.76
		IIC	-0.8805	**0.8849**	0.61
		IC_{hd}	-0.8712	**0.8835**	0.49
	R&G	IC	-0.8689	**0.8803**	0.38
		IIC	-0.8848	**0.8949**	0.41
		IC_{hd}	-0.8747	**0.8928**	0.38

1. Sim'_{res} always obtained better correlations values than its original formulation when our model is applied. The combination of the modified expression with the IC_{hd} approach improved the corpus-dependent metric while it remains as competitive as with IIC metric. For the larger dataset (R&G) this measure showed some stability in the ratio between the semantic differences and the semantic commonalities, close to the value $\frac{\beta}{\alpha} = 0.6$. This trend was not observed for the M&C dataset, probably due to the small number of pairs in the dataset. This result shows a higher importance for the commonalities when the larger dataset is used. For the M&C dataset the semantic differences played a more important role than the semantic commonalities, however the correlation with human judgments were higher for the R&G dataset.

4.4.3 Significance Analysis of the Results

In this section we asses the quality of the results by doing a significance test to the obtained correlations (using a similar method to the one employed in Section 3.4.3, Chapter 3). With the same datasets as in previous chapter (M&C and R&G) we choose the "difference between two dependent correlations test." This test allows to check if the correlation obtained using the Menendez-Ichise model is significantly different than the correlation obtained for the original metric and for the best reported result ($Sim_{p\&s}$).

Since the difference between the means of samples from two normal distributions is itself distributed normally, the T-distribution can be used to examine whether that difference can reasonably be supposed to be zero. The difference between two dependent correlations test assumes for the *null hypothesis* there is correlation between the tested correlations, despite the redundancy. While the *alternative hypothesis* is they are not correlated.

Table A.3, included in Appendix A, contains the upper critical values of the Student's t-distribution. The most commonly used significance level is $\alpha = 0.05$. For a two-sided test, we computed the percent point function at $\alpha/2$ (0.025). If the absolute value of the test statistic is greater than the upper critical value,then we reject the null hypothesis.

The expression for computing the t-values for the difference of correlations between datasets X_1 and X_2 is the one we previously introduced in Equation 3.22.

Returning to our experiments we analyze the results for Sim'_{res} (Equation 4.17), Sim'_{length} (Equation 4.14), Sim'_{lch} (Equation 4.16) and $Sim'_{j\&c}$ (Equation 4.18) versus their original version (without using the Menendez-Ichise model). We also compare our best result with the best reported result in the field ($Sim_{p\&s}$) [67].

Table 4.15 shows in detail the correlation values of the original Resnik's similarity (Sim_{res}) using the traditional approach for computing the information content (RES-IC), the correlation values of the original Pirró & Seco's similarity ($Sim_{p\&s}$) using the IIC approach of information content (PS-IIC) and the correlation values of the modified version of Resnik's similarity (after applied our model, Sim'_{res}) using the IIC and the IC_{hd} approaches of information content (RES'-IIC and RES'-IC_{hd}). The application of

the Menendez-Ichise model achieved higher values of correlations, but now our goal is to analyze the significance of those values when compared to the correlation of the original version of the similarity (RES-IC) and when compared with the best result published in the literature (PS-IIC).

Table 4.15: *Correlation values for Sim'_{res} similarity in the R&G dataset using different IC metrics compared with $Sim_{p\&s}$ similarity.*

	Sim_{res}	Sim'_{res}		$Sim_{p\&s}$
Correlation values	**RES -** IC	**RES' -** IIC	**RES' -** IC_{hd}	**PS-**IIC
P&S ratings	0.8677	0.8949	0.8928	0.8944
RES-IC	-	0.9652	0.9564	0.9603
RES'-IIC	-	-	0.9957	0.9994
RES'-IC_{hd}	-	-	-	0.9994

Table 4.16 shows the results of the significance tests for Sim'_{res} similarity for a sample size (N) of 28 and 65. When those results are compare with the upper critical values of Student's T-distribution for (N-3) degrees of freedom, see Table A.3 in Appendix A, we arrive to the following conclusions about the significance of the obtained correlations for the modified version of Resnik's similarity Sim'_{res}:

Table 4.16: *Significance values for Sim'_{res} in the R&G dataset using IIC metric when compare with the original similarity (RES-IC) and the P&S similarity ($Sim_{p\&s}$) using the IIC metric.*

Values of the t-statistics		
Sample size	**RES-**IC **vs. RES'-**IIC	**RES'-**IIC **vs. PS-**IIC
28	1.1559	0.1689
65	2.8203	0.2660

- The application of the Menendez-Ichise model allowed us to obtain a significantly better similarity (RES'-IIC) when compare with it's original version (RES-IC) for a sample size of 65 (for the smaller dataset (M&C) the results were not significant).

- The results from the combination of the modified version of Resnik's similarity (Sim'_{res}) with the intrinsic information content (RES'-IIC) were not significant different from the Pirró & Seco similarity ($Sim_{p\&s}$) (also using the intrinsic information content IIC) for none of the datasets.

- However as shown in Table 4.17 the application of our model allowed to obtain the smallest values for the standard deviation when compared with the other two similarities.

Table 4.17: *Descriptive analysis for Sim'_{res} using IIC metric, Sim_{res} using IC metric and the P&S similarity ($Sim_{p\&s}$) using the IIC metric.*

		Sim_{res}	Sim'_{res}		$Sim_{p\&s}$
	P&S ratings	**RES-**IC	**RES'-**IIC	**RES'-**IC_{hd}	**PS-**IIC
Min	0.3929	0.0000	-0.7244	-0.7116	-1.8154
Max	3.4300	12.1630	0.6000	0.6000	1.0000
Average	1.5398	4.5970	-0.1536	-0.1381	-0.5717
Median	1.1663	2.3335	-0.3662	-0.3050	-1.0064
Mode	-	1.2900	0.6000	0.6000	1.0000
STDEV	1.0027	3.9622	0.4547	0.3960	0.9748
STDER	-	0.5023	0.4510	0.4555	0.4520

The same analysis of significance was done for the other similarity measures where the Menendez-Ichise model was applied (Sim'_{length}, and Sim'_{lch}. For none of them the improvements in the results were significantly different as shown in Table C.1, Table C.2, Table C.3, Table C.4, Table C.5 and Table C.6, in Appendix C. In the case of the $Sim'_{j\&c}$, it's modified expression was the same as the obtained for Sim'_{res} but again the results although better, they were not significantly different.

4.5 General Discussion

Summarizing the results of experiments we can state the application of Menendez-Ichise model showed positive results for the Sim'_{length}, Sim'_{lch}, $Sim'_{j\&c}$ and Sim'_{res} measures which obtained higher values of correlation than their original expressions when the semantic differences between the concepts were also taken into consideration.

The Sim'_{wup} and the Sim'_{lin} measures were not affected by the application of our model. In fact, due to the design of those measures a ratio approach rather than the contrast approach would possibly lead to better results. The experiments also showed the similarity values obtained by each modified similarity had a higher stability (lower standard deviation).

The use of the IC_{hd} approach for the node-based measures always showed better results than the corpus-dependent approach while remaining as competitive as the IIC metric, in the case of Sim'_{lin} it allowed to obtain the highest correlation value. Sim'_{wup} measure remain the same as its original version. All node-based similarity measures were superior to the edge-based ones. Curiously, the results of Sim'_{length} were better than those for Sim'_{lch} for the M&C dataset, despite the simplicity of its model.

The experiments also suggested a larger dataset could be helpful for estimating the best ratio of the importance between the semantic differences and the semantic commonalities. However based of the results of the largest dataset (R&G) for node-based similarity measures the semantic differences had a higher importance than the semantic commonalities in the final result, but the opposed was the case for the edge-based similarities.

4.6 Chapter Conclusion

In the present chapter we have introduced new ideas in the computation of WordNet-based semantic similarity measures. We developed five new measures which are modifications of traditional WordNet-based semantic similarity metrics. Supported by a featured-based theory, they incorporate the idea of semantic differences between concepts into the similarity computation process. The experimental results showed that, four of the measures outperformed their classical or original version, while the other measure performed the same as its classical version. These results demonstrate the strengths and positive effects of including concepts semantic differences and its combination with the proposed information content metric during their semantic similarity computations. The extended corpora independent approach generated the highest value for one of the node-based measure, and in general it improved the results of the corpus-dependent model while remained as competitive as the intrinsic information content approach.

This research focus on WordNet-based semantic similarity measures. The studied similarity measures use the hyponymy relation, also known as the "IS-A" relation, for computing the similarity between two concepts. Despite of the fact that about 80% of the

relationships in the WordNet taxonomy are "IS-A" relationships, it is a shortcoming of those measures not to consider other types of relations. The term "semantic relatedness" refers to several types of lexical relationships, including synonymy, meronymy, antonymy, as well as any other unsystematic relationships, i.e. functional relationships. The application of our approach to semantic relatedness measures is possible and it remains as an open area of research.

As future work, we would like to enlarge the words' pairs dataset. This could help us to estimate the ratio between semantic differences and semantic commonalities. We also would like to apply some machine learning methods to estimate the best ratio between differences and commonalities, and finally to apply our developed measures to a real problem.

Chapter 5

Conclusion and Future Work

The final chapter of this thesis aims to summarize the main outcomes and possible future developments of our research efforts to provide an efficient corpus-independent information content metric as well as to improve existing WordNet based semantic similarity measures. The following sections present the contributions of this research (Section 5.1), limitations and future work (Section 5.2).

5.1 Contributions

The main contributions of the thesis are:

A novel model for semantic similarity computation. We showed that a featured based model of similarity, where semantic differences and semantic commonalities are both considered, can be applied to word pairs comparison. We demonstrated the model application by obtaining 5 new semantic similarity measures.

Five new semantic similarity measures. After applying the Menendez-Ichise model to the traditional WordNet based semantic similarity measures we obtained five new measures. We showed four of this similarity measures outperformed their classical version while the last one performed the same as its' classical version.

A new corpora independent information content metric. We showed an analysis of taxonomic properties in corpus independent metrics. The application of this analysis allowed us to obtain a new corpora independent information content metric which generated the highest value of accuracy among the corpora dependent and the corpora independent metrics.

5.2 Discussion

In the current section, we will outline a set of problems and open questions that we had to leave out for time limitations with the hope to come back to them at a later point in the near future. We will roughly sketch directions for possible solutions of some of these problems, but most of them will be left simply as a list of related, unsolved issues.

5.2.1 Limitations

Section 4.3 introduced to our proposed model for semantic similarity computation. There is a generalization limitation in the application of our model because we only considered the *IS-A* relations of WordNet (the hierarchy taxonomy). Another limitation is related with how the WordNet taxonomy was built, which forced us to assume symmetry in the importance of the relations.

Section 3.3 introduced a corpus-independent information content metric. In the analysis of the taxonomic properties we focused in the depth of a concept in the taxonomy ($depth(c)$). So, when taxonomies (or sub taxonomies) with low depth but a wide spread in its structure are used, the proposed information content metric (IC_{hd}) will probably be affected by the structure of the taxonomy.

5.2.2 Future Work

I have started this work motivated by the ontology matching and the interoperability problem of the semantic web. In the future I would like to contribute toward an application of this work in the previous mentioned areas. Here I will be discussing about my ongoing work and a few possible work that might be undertaken in the near future.

- For some domains when new knowledge is inserted, rather than growing up as a new level, they grow up to the sides as new sibling instead of a child. From this perspective new features related to the *width* of the taxonomy and the *number of siblings* a concept have, should be considered for computing the information content of a concept. In the future we would like to extend the corpus-independent information content IC_{hd} we developed so it also consider the *number of siblings* a concept have as well as the *width* of the taxonomy. Probably the ratio between the maximum width and the maximum depth of the taxonomy should be analyzed.

- Since the current M&C and R&G word pairs datasets are limited to very small amount of word pairs it would be useful to convey an study to increase the human judgment about the semantic similarity and relatedness of a higher amount of word pairs.

- Apply the Menendez-Ichise Model to semantic relatedness measures, where not only *IS-A* relations were considered.

- To extend the Menendez-Ichise Model to consider asymmetric importance of the relations linking the concepts in the taxonomy.

- To study how to assign weights to each of the relations present in WordNet.

5.3 Summary

This thesis has developed and apply a model of semantic similarity computation for word pair comparison. This model consider the semantic commonalities and the semantic differences as the core of its approach. By applying the model five new WordNet-based semantic similarity measures for word pair comparison were created. Four of this semantic similarity measures obtained higher values of correlation with human judgment than their original expressions, while the fifth one remained as competitive as their original version. We also study WordNet taxonomic properties to extend a corpus-independent information content metric. The application of this new metric in one of the previously developed node-based semantic similarity allowed us to obtain the highest value of correlation with respect to human judgment. This thesis provides a general and extensible approach of semantic similarity computation for word pair comparison.

Bibliography

[1] S. Auwatanamongkol. Inexact graph matching using a genetic algorithm for image recognition. *Pattern Recognition Letters*, 28(12):1428–1437, 2007.

[2] S. Banerjee and T. Pedersen. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proc. 18th International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, 2003.

[3] M. Batet, D. Sanchez, A. Valls, and K. Gibert. Exploiting taxonomical knowledge to compute semantic similarity: An evaluation in the biomedical domain. In N. Garca-Pedrajas, F. Herrera, C. Fyfe, J. Bentez, and M. Ali, editors, *Trends in Applied Intelligent Systems*, volume 6096 of *Lecture Notes in Computer Science*, pages 274–283. Springer Berlin / Heidelberg, 2010.

[4] M. Batet, D. Snchez, and A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1):118 – 125, 2011.

[5] C. Batini, M. Lenzerini, and S. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4):323–364, 1986.

[6] E. Bengoetxea. *Inexact Graph Matching Using Estimation of Distribution Algorithms*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, Dec. 2002.

[7] T. Berners-Lee and M. Fischetti. *Weaving the Web: The original design and ultimate destiny of the World Wide Web, by its inventor*. Harper San Francisco, 1999.

[8] D. Bollegala, Y. Matsuo, and M. Ishizuka. A web search engine-based approach to measure semantic similarity between words. *IEEE Transactions on Knowledge and Data Engineering*, 23:977–990, 2011.

[9] A. Budanitsky and G. Hirst. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Proc. Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Comnputational Linguistics*, pages 29–34, 2001.

[10] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47, Mar. 2006.

[11] H. Bunke. Graph matching: Theoretical foundations, algorithms, and applications. *Proc. Vision Interface*, 23(2):82–88, 2000.

[12] W. W. Cohen, P. Ravikumar, S. Fienberg, and K. Rivard. SecondString Project. http://secondstring.sourceforge.net/.

[13] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *International Joint Conference on Artificial Intelligence, Workshop on Information Integration on the Web*, pages 73–78, 2003.

[14] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, Nov. 1975.

[15] V. Cross and X. Yu. Investigating ontological similarity theoretically with fuzzy set theory, information content, and tversky similarity and empirically with the gene ontology. In S. Benferhat and J. Grant, editors, *Proc. 5th International Conference Scalable Uncertainty Management*, volume 6929 of *Lecture Notes in Computer Science*, pages 387–400. Springer Berlin / Heidelberg, 2011.

[16] J. David, J. Euzenat, F. Scharffe, and C. T. dos Santos. The alignment api 4.0. *Semantic Web Journal*, 2(1):3–10, 2011.

[17] J. David, F. Guillet, and H. Briand. Association Rule Ontology Matching Approach. *International Journal on Semantic Web and Information Systems*, 3(2), 2007.

[18] M. Deza and E. Deza. *Encyclopedia of Distances*. Springer-Verlag, 2009.

[19] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proc. 11th International Conference on World Wide Web*, WWW '02, pages 662–673, New York, NY, USA, 2002. ACM.

[20] H. Dong, F. K. Hussain, and E. Chang. A context-aware semantic similarity model for ontology environments. *Concurrency and Computation: Practice and Experience*, 23(5):505–524, 2011.

[21] D. Estival, C. Nowak, and A. Zschorn. Towards ontology-based natural language processing. In *Proc. Workshop on NLP and XML: RDF/RDFS and OWL in Language Technology*, pages 59–66, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[22] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[23] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. Stephens. The Semantic Web in Action. *Scientific American*, 297:90–97, Nov. 2007.

[24] C. Fellbaum, editor. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1st edition, 1998.

[25] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening Ontologies with DOLCE. In *Proc. 13th International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 166–181, London, UK, 2002. Springer-Verlag.

[26] D. Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[27] S. J. Green. Building Hypertext Links By Computing Semantic Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 11(5):713–730, 1999.

[28] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43(5-6):907–928, 1995.

[29] J. Hai and C. Hanhua. SemreX: Efficient Search in a Semantic Overlay for Literature Retrieval. *Future Generation Computer Systems*, 11(6):475–488, 2008.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[31] A. Hameed, A. Preece, and D. Sleema. Ontology reconciliation. In S. Staab and R. Studer, editors, *Handbook of ontologies*, International handbooks on information systems, chapter 12, pages 231–250. Springer Verlag, 2003.

[32] G. Hirst and D. St-Onge. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database (Language, Speechand Communication)*, chapter 13, pages 305–332. The MIT Press, 1998.

[33] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. E. Milios. Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems*, 2(3):55–73, 2006.

[34] B. Hu, S. Dasmahapatra, and P. Lewis. Semantic metrics. *International Journal of Metadata, Semantics and Ontologies*, 2(4):242–258, 2007.

[35] S. Imai. Pattern similarity and cognitive transformations. *Acta Psychologica*, 41(16):433–447, 1977.

[36] K. Janowicz. Sim-DL: Towards a semantic similarity measurement theory for the description logic $\mathcal{ALCNR}$ in geographic information retrieval. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems*, volume 4278 of *Lecture Notes in Computer Sciences*, pages 1681–1692. Springer, 2006.

[37] K. Janowicz, B. Adams, and Raub. Semantic referencing - determining context weights for similarity measurement. In *Proc. 6th International Conference on Geographic Information Science*, pages 70–84, Berlin, Heidelberg, 2010. Springer-Verlag.

[38] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proc. International Conference on Research in Computational Linguistic*, pages 19–33, 1997.

[39] Y. Kalfoglou and M. Schorlemmer. If-map: an ontology mapping method based on information flow theory. *Journal on Data Semantics*, 1(1):98–127, Oct. 2003.

[40] C. Keßler. Similarity measurement in context. In *Proc. 6th International and Interdisciplinary Conference on Modeling and Using Context*, pages 277–290, Berlin, Heidelberg, 2007. Springer-Verlag.

[41] C. Keßler, M. Raubal, and K. Janowicz. The effect of context on semantic similarity measurement. In *Proc. Confederated International Conference on On the Move to Meaningful Internet Systems - Volume Part II*, pages 1274–1284, Berlin, Heidelberg, 2007. Springer-Verlag.

[42] M. Klein. Combining and relating ontologies: an analysis of problems and solutions. In *Proc. 17th International Joint Conference on Artificial Intelligence, Workshop on Ontologies and Information Sharing*, pages 16–39, 2001.

[43] G. Kondrak. N-gram similarity and distance. In *Proc. International Conference on String Processing and Information Retrieval*, pages 115–126, 2005.

[44] M. S. Lacher and G. Groh. Facilitating the exchange of explicit knowledge through ontology mappings. In *Proc. 14th International FLAIRS Conference*, pages 305–309. AAAI Press, 2001.

[45] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In *WordNet: A Lexical Reference System and its Application*, pages 265–283, 1998.

[46] H. Li, Y. Tian, and Q. Cai. Improvement of semantic similarity algorithm based on wordnet. In *6th IEEE Conference on Industrial Electronics and Applications*, pages 564 –567, June 2011.

[47] Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, 2003.

[48] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conference on Machine Learning*, pages 296–304, 1998.

[49] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. *The International Journal on Very Large Data Bases*, 10:49–58, 2001.

[50] A. Maedche and S. Staab. Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[51] P. McCorduck. *Machines Who Think*. Natick, MA: A K Peters, Ltd., 2nd edition, 2004.

[52] S. McDonald. A context-based model of semantic similarity. *ACM Transactions on Programming Languages and Systems*, 15(5):795–825, Nov. 1997.

[53] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *Proc. 17th International Conference on Principles of Knowledge Representation and Reasoning*, pages 483–493, Apr. 2000.

[54] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. Mooney. Active Learning for Probability Estimation using Jensen-Shannon Divergence. In *Proc. 16th European Conference on Machine Learning*, pages 268–279, 2005.

[55] R. E. Menéndez-Mora and R. Ichise. Effect of Semantic Differences in WordNet-Based Similarity Measures. In *Proc. 23rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems*, volume LNAI 6097, pages 545–554, 2010.

[56] R. E. Menéndez-Mora and R. Ichise. The Role of Taxonomy Properties in Information Content Metrics. In *Proc. International Symposium on Matching and Meaning 2010*, pages 22–26, 2010.

[57] R. E. Menéndez-Mora and R. Ichise. Toward Simulating the Human Way of Comparing Concepts. *IEICE TRANSACTIONS on Information and Systems*, E94-D(7):1419–1429, July 2011.

[58] G. A. Miller and W. G. Charles. Contextual correlates of semantic synonymy. *Languages and Cognitive Processes*, 6(1):1–28, 1991.

[59] P. Mitra and G. Wiederhold. An ontology composition algebra. In S. Staab and R. Studer, editors, *Handbook on Ontologies*. Springer, 2004.

[60] R. Mizoguchi. Tutorial on ontological engineering part 2: Ontology development, tools and languages. *Journal New Generation Computing*, 22(1):61–96, 2004.

[61] S. Mohammad and G. Hirst. Distributional measures as proxies for semantic relatedness. *Kluwer Academic Publishers*, pages 1–47, 2005.

[62] J. Morris and G. Hirst. Non-classical lexical semantic relations. In *Proc. HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 46–51, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[63] N. F. Noy and M. A. Musen. Anchor-prompt: Using non-local context for semantic matching. In *Proc. Workshop on Ontologies and Information Sharing*, pages 63–70, 2001.

[64] E. Ovchinnikova and K.-U. Kuehnberger. Automatic ontology extension: Resolving inconsistencies. *GLDV-Journal for Computational Linguistics and Language Technology*, 22(2):19–33, 2007.

[65] E. Ovchinnikova, T. Wandmacher, and K.-U. Kuehnberger. Solving terminological inconsistency problems in ontology design. *IBIS-Interoperability in Business Information Systems*, 2:65–80, 2007.

[66] A. Pease, I. Niles, and J. Li. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, pages 1–4, 2002.

[67] G. Pirró. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering*, 68(11):1289–1308, 2009.

[68] G. Pirró and J. Euzenat. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, and B. Glimm, editors, *Proc. 9th International Semantic Web Conference*, volume 6496 of *Lecture Notes in Computer Sciences*, pages 615–630, Springer Berlin / Heidelberg, Nov. 2010.

[69] G. Pirró and J. Euzenat. A semantic similarity framework exploiting multiple parts-of speech. In R. Meersman, T. Dillon, and P. Herrero, editors, *On the Move to Meaningful Internet Systems, OTM 2010*, volume 6427 of *Lecture Notes in Computer Science*, pages 1118–1125. Springer Berlin / Heidelberg, 2010.

[70] G. Pirró and N. Seco. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. In *Proc. On the Move to Meaningful Internet Systems: OTM 2008*, pages 1271–1288, 2008.

[71] J. Piskorski, M. Sydow, and K. Wieloch. Comparison of String Distance Metrics for Lemmatisation of Named Entities in Polish. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society*, volume 5603 of *Lecture Notes in Computer Science*, pages 413–427. Springer Berlin / Heidelberg, 2009.

[72] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[73] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The International Journal on Very Large Data Bases*, 10(4):334–350, 2001.

[74] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Aug. 1995.

[75] P. Resnik. Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. *Artificial Intelligence Research*, 11:95–130, 1999.

[76] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of ACM*, 8(10):627–633, 1965.

[77] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Prentice Hall, 2003.

[78] D. Sánchez and M. Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, 44(5):749 – 759, 2011.

[79] D. Sánchez, M. Batet, and D. Isern. Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297 – 303, 2011.

[80] D. Sánchez, M. Batet, and A. Valls. Web-based semantic similarity: An evaluation in the biomedical domain. *International Journal of Software and Informatics*, 4(1):39–52, 2010.

[81] K. Saruladha, G. Aghila, and A. Bhuvaneswary. Information content based semantic similarity approaches for multiple biomedical ontologies. In A. Abraham, J. Lloret Mauri, J. F. Buford, J. Suzuki, and S. M. Thampi, editors, *Advances in Computing and Communications*, volume 191 of *Communications in Computer and Information Science*, pages 327–336. Springer Berlin Heidelberg, 2011.

[82] A. D. Scriver. Semantic distance in wordnet: A simplified and improved measure of semantic relatedness. Master's thesis, University of Waterloo, 2006.

[83] N. Seco. Computational Models of Similarity in Lexical Ontologies. Master's thesis, University College Dublin, 2005.

[84] N. Seco, T. Veale, and J. Hayes. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proc. European Conference on Artificial Intelligence*, pages 1089–1090, 2004.

[85] A. P. Sheth and J. A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22:183–236, 1990.

[86] L. Sjoberg. A cognitive theory of similarity. *Goteborg Psychological Reports*, 2(10), 1972.

[87] J. F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA, 1st edition, Aug. 1999.

[88] R. K. Srihari, Z. Zhang, A. Rao, H. Baird, and F. Chen. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2–3):245–275, 2000.

[89] S. Staab and R. Studer, editors. *Handbook of Ontologies*. International Handbooks on Information Systems. Springer, 2009.

[90] G. Stoilos, G. Stamou, and S. Kollias. A String Metric for Ontology Alignment. In *Proc. 4th International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Sciences*, pages 624–637, 2005.

[91] U. Straccia and R. Troncy. oMAP: Combining classifiers for aligning automatically OWL ontologies. In *6th International Conference on Web Information Systems Engineering*, volume 3806 of *Lecture Notes in Computer Sciences*, pages 133–147. Springer, 2005.

[92] G. Stumme and A. Maedche. Fca-merge: Bottom-up merging of ontologies. In *Proc. International Joint Conference on Artificial intelligence*, pages 225–230, 2001.

[93] X. Su. A text categorization perspective for ontology mapping. Technical report, Department of Computer and Information Science. Norwegian University of Science and Technology, 2002.

[94] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proc. 2nd International Conference on Information and Knowledge Management*, pages 67–74, 1993.

[95] A. Thor, T. Kirsten, and E. Rahm. Instance-based matching of hierarchical ontologies. In *Proc. of the 12th Conference on Database Systems for Business, Technology and Web (BTW)*, pages 1–13, 2007.

[96] K. Todorov. *Ontology Matching by Combining Instance-Based Concept Similarity Measures with Structure*. PhD thesis, University of Osnabrück, Oct. 2009.

[97] W. S. Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393, 1965.

[98] C. Tsinaraki, Y. Velegrakis, N. Kiyavitskaya, and J. Mylopoulos. A context-based model for the intrerpretation of polysemous terms. In R. Meersman, T. Dillon, and P. Herrero, editors, *On the Move to Meaningful Internet Systems, OTM 2010*, volume 6427 of *Lecture Notes in Computer Sciences*, pages 939–956. Springer Berlin Heidelberg, 2010.

[99] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[100] S. Wang, G. Englebienne, and S. Schlobach. Learning concept mappings from instance similarity. In H. Berlin, editor, *Proc. 7th International Conference on The Semantic Web*, pages 339–355. Springer-Verlag, 2008.

[101] C. Wartena and R. Brussee. Instanced-based mapping between thesauri and folksonomies. In H. Berlin, editor, *Proc. 7th International Conference on The Semantic Web*, pages 356–370. Springer-Verlag, 2008.

[102] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. In *Proc. 32nd Annual Meeting of the Association for Computational Linguistic*, pages 133–138, 1994.

[103] D. Yang and D. M. W. Powers. Measuring semantic similarity in the taxonomy of wordnet. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38*, ACSC '05, pages 315–322, Darlinghurst, Australia, 2005. Australian Computer Society, Inc.

[104] Z. Zhang, A. L. Gentile, and F. Ciravegna. Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 991–1002, 2011.

[105] W. Zhou, H. Wang, J. Chao, W. Zhang, and Y. Yu. Loddo: Using linked open data description overlap to measure semantic relatedness between named entities. In *Proc. of the Joint International Semantic Technology Conference*, pages 266–281, Dec. 2011.

[106] Z. Zhou, Y. Wang, and J. Gu. A new model of information content for semantic similarity in wordnet. In *Proc. 2nd International Conference on Future Generation Communication and Networking Symposia*, volume 3, pages 85 –89, 2008.

Appendix A

General Details for Experiments

Table A.1: *Human judgments in the Pirró and Seco experiment for M&C word pairs dataset*

	Pair	P&S ratings
1	gem jewel	3.2217
2	midday noon	3.247
3	automobile car	3.421
4	boy lad	3.0256
5	implement tool	2.6444
6	coast shore	3.0257
7	journey voyage	2.9452
8	magician wizard	2.8845
9	furnace stove	2.3248
10	asylum madhouse	2.6581
11	brother monk	1.9971
12	food fruit	1.7569
13	bird cock	1.6606
14	bird crane	1.6838
15	brother lad	1.5249
16	crane implement	1.3563
17	car journey	1.176
18	coast hill	0.9611
19	food rooster	1.0477
20	forest graveyard	0.8008
21	lad wizard	0.7519
22	monk oracle	1.1663
23	coast forest	0.7325
24	monk slave	0.6948
25	glass magician	0.5093
26	noon string	0.4425
27	rooster voyage	0.4211
28	cord smile	0.476

Table A.2: *Human judgments in the Pirró and Seco experiment for R&G word pairs dataset*

	Pair	P&S ratings
1	cemetery graveyard	3.43
2	automobile car	3.421
3	midday noon	3.247
4	gem jewel	3.2217
5	cock rooster	3.1431
6	cushion pillow	3.1343
7	coast shore	3.0257
8	boy lad	3.0256
9	forest woodland	2.9749
10	journey voyage	2.9452
11	magician wizard	2.8845
12	grin smile	2.7131
13	autograph signature	2.6759
14	asylum madhouse	2.6581
15	implement tool	2.6444
16	cord string	2.6137
17	serf slave	2.5536
18	hill mound	2.5345
19	glass tumbler	2.5036
20	furnace stove	2.3248
21	oracle sage	2.2164
22	brother monk	1.9971
23	sage wizard	1.893
24	food fruit	1.7569
25	bird crane	1.6838
26	bird cock	1.6606
27	magician oracle	1.5898
28	brother lad	1.5249
29	crane implement	1.3563
30	crane rooster	1.2712
31	cemetery mound	1.253
32	car journey	1.176
33	monk oracle	1.1663
34	glass jewel	1.1459
35	furnace implement	1.1114
36	hill woodland	1.0655
37	food rooster	1.0477
38	coast hill	0.9611
39	shore voyage	0.931
40	bird woodland	0.8779
41	forest graveyard	0.8008
42	shore woodland	0.7563
43	mound shore	0.7531
44	lad wizard	0.7519

Continued on Next Page...

Table A.2 – Continued

	Pair	P&S ratings
45	coast forest	0.7325
46	asylum monk	0.721
47	cemetery woodland	0.6969
48	monk slave	0.6948
49	asylum cemetery	0.6721
50	boy rooster	0.6432
51	grin lad	0.5886
52	boy sage	0.5847
53	automobile cushion	0.5843
54	cushion jewel	0.5819
55	graveyard madhouse	0.5615
56	grin implement	0.5147
57	glass magician	0.5093
58	mound stove	0.4943
59	cord smile	0.476
60	automobile wizard	0.4606
61	autograph shore	0.45
62	noon string	0.4425
63	fruit furnace	0.4384
64	rooster voyage	0.4211
65	asylum fruit	0.3929

Table A.3: *Upper critical values of Student's T-distribution with* $\boldsymbol{v}$ *degrees of freedom*

	Probability of exceeding the critical value					
v	0.1	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.92	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.44	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.86	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.25	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.35	1.771	2.16	2.65	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.12	2.583	2.921	3.686
17	1.333	1.74	2.11	2.567	2.898	3.646
18	1.33	1.734	2.101	2.552	2.878	3.61
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.08	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.5	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	***1.316***	***1.708***	***2.06***	***2.485***	***2.787***	***3.45***
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.31	1.697	2.042	2.457	2.75	3.385
			...			
60	1.296	1.671	2	2.39	2.66	3.232
61	1.296	1.67	2	2.389	2.659	3.229
62	***1.295***	***1.67***	***1.999***	***2.388***	***2.657***	***3.227***
63	1.295	1.669	1.998	2.387	2.656	3.225
64	1.295	1.669	1.998	2.386	2.655	3.223
65	1.295	1.669	1.997	2.385	2.654	3.22
66	1.295	1.668	1.997	2.384	2.652	3.218
67	1.294	1.668	1.996	2.383	2.651	3.216
68	1.294	1.668	1.995	2.382	2.65	3.214
69	1.294	1.667	1.995	2.382	2.649	3.213
70	1.294	1.667	1.994	2.381	2.648	3.211

Table A.4: *Results of the original similarities for Miller and Charles word pairs dataset*

	Length	Sim_{wup}	Sim_{lch}	Sim_{lin}	Sim_{res}	$Sim_{J\&C}$	$Sim_{P\&S}$ IIC
1	1.0000	1.0000	3.6889	1.0000	12.0677	0.0000	1.0000
2	1.0000	1.0000	3.6889	1.0000	11.0644	0.0000	1.0000
3	1.0000	1.0000	3.6889	1.0000	7.5914	0.0000	1.0000
4	0.5000	0.9333	2.9957	0.8306	8.3995	3.4271	0.4796
5	0.5000	0.9412	2.9957	0.9472	5.8774	0.6558	0.3526
6	0.5000	0.9231	2.9957	0.9632	9.4157	0.7191	0.7747
7	0.5000	0.9565	2.9957	0.8093	7.1100	3.3509	0.6178
8	1.0000	1.0000	3.6889	1.0000	11.9807	0.0000	1.0000
9	0.1000	0.5714	1.3863	0.2281	2.3058	15.6028	-0.9979
10	0.5000	0.9565	2.9957	0.8556	9.4752	3.1987	0.8775
11	0.5000	0.9565	2.9957	0.9864	9.2616	0.2552	0.7550
12	0.1000	0.4706	1.3863	0.1610	1.5928	16.6022	-1.4213
13	0.5000	0.9565	2.9957	0.7739	7.6778	4.4853	-0.1354
14	0.2500	0.8800	2.3026	0.7478	7.6778	5.1784	-0.1354
15	0.2000	0.7143	2.0794	0.2552	2.3335	13.7908	-1.1627
16	0.2000	0.7778	2.0794	0.3591	3.2577	11.6304	-0.5783
17	0.2000	0.1905	0.7985	0.0000	0.0000	14.4174	-1.2999
18	0.0625	0.7143	2.0794	0.5991	5.8847	7.8753	-0.0448
19	0.1111	0.2857	0.9163	0.0919	0.8018	15.8389	-1.0972
20	0.2000	0.5000	1.4917	0.1234	1.2900	18.3211	-1.5097
21	0.1250	0.7143	2.0794	0.2551	2.3335	13.6287	-1.2239
22	0.1667	0.5882	1.6094	0.2257	2.3335	16.0156	-1.1014
23	0.2000	0.6154	1.8971	0.1306	1.2900	17.1684	-1.4288
24	0.1250	0.7143	2.0794	0.2543	2.3335	13.6848	-1.0120
25	0.0833	0.5333	1.6094	0.2142	2.2826	17.0388	-1.2902
26	0.0909	0.3529	1.2040	0.0662	0.5962	19.4623	-1.8003
27	0.0556	0.1481	0.5108	0.0000	0.0000	23.0733	-1.8154
28	0.0417	0.3750	1.2910	0.0540	0.5962	20.9100	-1.7033

Appendix B

Details of Experimental Results. A Corpus Independent Information Content Metric

Table B.1, shows the results of computing Lin's similarity (Sim_{lin}) for each word pair in M&C dataset. The first column (P&S ratings) represents the human judgments of the 101 participants in the P&S experiment for the M&C dataset. The rest of the columns show the results of computing Lin's similarity whit different IC metrics (IC, IIC and IC_{hd}).

Table B.1: *Results of Sim_{lin} measure for each word pair in the M&C dataset using different IC metrics.*

P&S ratings	Sim_{lin} - IC	Sim_{lin} - IIC	Sim_{lin} - IC_{hd}
3.2217	1.0000	1.0000	1.0000
3.2470	1.0000	1.0000	1.0000
3.4210	1.0000	1.0000	1.0000
3.0256	0.8306	0.8504	0.7614
2.6444	0.9472	0.9277	0.9272
3.0257	0.9632	0.9752	0.9690
2.9452	0.8093	0.9169	0.9006
2.8845	1.0000	1.0000	1.0000
2.3248	0.2281	0.2355	0.2476
2.6581	0.8556	0.9684	0.9089
1.9971	0.9864	0.9347	0.8642
1.7569	0.1610	0.1779	0.1663
1.6606	0.7739	0.5993	0.5974
1.6838	0.7478	0.5993	0.6020
1.5249	0.2529	0.2399	0.2444
1.3563	0.3591	0.3690	0.3877
1.1760	0.0000	0.0132	0.0239
0.9611	0.5991	0.6488	0.6396
1.0477	0.0919	0.1000	0.1057
0.8008	0.1234	0.1014	0.1125
0.7519	0.2551	0.2321	0.2349
1.1663	0.2257	0.2483	0.2526
0.7325	0.1306	0.1062	0.1211
0.6948	0.2543	0.2616	0.2701
0.5093	0.2113	0.2147	0.1984
0.4425	0.0577	0.0666	0.0632
0.4211	0.0000	0.0095	0.0167
0.4760	0.0540	0.0699	0.0714

Table B.2, shows the results of computing Resnik's similarity (Sim_{res}) for each word pair in M&C dataset. The first column (P&S ratings) represents the human judgments of the 101 participants in the P&S experiment for the M&C dataset. The rest of the columns show the results of computing Resnik's similarity whit different IC metrics (IC, IIC and IC_{hd}).

Table B.2: *Results of Sim_{res} measure for each word pair in the M&C dataset using different IC metrics.*

P&S ratings	Sim_{res} - IC	Sim_{res} - IIC	Sim_{res} - IC_{hd}
3.2217	12.0677	1.0000	1.0000
3.2470	11.0644	1.0000	1.0000
3.4210	7.5914	0.6718	0.5899
3.0256	8.3995	0.7398	0.6147
2.6444	5.8774	0.4177	0.3626
3.0257	9.4157	0.8162	0.6753
2.9452	7.1100	0.7547	0.6190
2.8845	11.9807	0.8162	0.6800
2.3248	2.3058	0.1817	0.1666
2.6581	9.4752	0.9387	0.8330
1.9971	9.2616	0.8775	0.7609
1.7569	1.5928	0.1725	0.1506
1.6606	7.6778	0.4017	0.3682
1.6838	7.6778	0.4017	0.3682
1.5249	2.3335	0.2179	0.1906
1.3563	3.2577	0.2390	0.2166
1.1760	0.0000	0.0088	0.0139
0.9611	5.8847	0.5431	0.4472
1.0477	0.8018	0.0645	0.0616
0.8008	1.2900	0.0902	0.0858
0.7519	2.3335	0.2179	0.1906
1.1663	2.3335	0.2179	0.1906
0.7325	1.2900	0.0902	0.0858
0.6948	2.3335	0.2179	0.1906
0.5093	2.2826	0.2042	0.1757
0.4425	0.5962	0.0666	0.0632
0.4211	0.0000	0.0088	0.0139
0.4760	0.5962	0.0666	0.0632

Table B.3, shows the results of computing Pirró and Seco's similarity ($Sim_{P\&S}$) for each word pair in M&C dataset. The first column (P&S ratings) represents the human judgments of the 101 participants in the P&S experiment for the M&C dataset. The rest of the columns show the results of computing Pirró and Seco's similarity whit different IC metrics (IC, IIC and IC_{hd}).

Table B.3: *Results of $Sim_{P\&S}$ measure for each word pair in the M&C dataset using different IC metrics.*

P&S ratings	$Sim_{P\&S}$ - IC	$Sim_{P\&S}$ - IIC	$Sim_{P\&S}$ - IC_{hd}
3.2217	12.0677	1.0000	1.0000
3.2470	11.0644	1.0000	1.0000
3.4210	7.5914	1.0000	1.0000
3.0256	4.9724	0.4796	0.2294
2.6444	5.2215	0.3526	0.3057
3.0257	8.6966	0.7747	0.6321
2.9452	3.7591	0.6178	0.4823
2.8845	11.9807	1.0000	1.0000
2.3248	-13.2970	-0.9979	-0.8459
2.6581	6.2765	0.8775	0.6660
1.9971	9.0063	0.7550	0.5217
1.7569	-15.0095	-1.4213	-1.3595
1.6606	3.1925	-0.1354	-0.1280
1.6838	2.4993	-0.1354	-0.1187
1.5249	-11.4573	-1.1627	-0.9881
1.3563	-8.3727	-0.5783	-0.4674
1.1760	-14.4174	-1.2999	-1.1168
0.9611	-1.9906	-0.0448	-0.0567
1.0477	-15.0372	-1.0972	-0.9807
0.8008	-17.0310	-1.5097	-1.2679
0.7519	-11.2951	-1.2239	-1.0510
1.1663	-13.6820	-1.1014	-0.9376
0.7325	-15.8784	-1.4288	-1.1594
0.6948	-11.3513	-1.0120	-0.8396
0.5093	-14.7561	-1.2902	-1.2435
0.4425	-18.8661	-1.8003	-1.8105
0.4211	-23.0733	-1.8154	-1.6179
0.4760	-20.3138	-1.7033	-1.5810

Table B.4, shows the results of computing Jiang & Conrath's similarity ($Sim_{J\&C}$) for each word pair in M&C dataset. The first column (P&S ratings) represents the human judgments of the 101 participants in the P&S experiment for the M&C dataset. The rest of the columns show the results of computing Jiang & Conrath's similarity whit different IC metrics (IC, IIC and IC_{hd}).

Table B.4: *Results of $Sim_{J\&C}$ measure for each word pair in the M&C dataset using different IC metrics.*

P&S ratings	$Sim_{J\&C}$ - IC	$Sim_{J\&C}$ - IIC	$Sim_{J\&C}$ - IC_{hd}
3.2217	12.0677	1.0000	1.0000
3.2470	11.0644	1.0000	1.0000
3.4210	7.5914	1.0000	1.0000
3.0256	4.9724	0.4796	0.2294
2.6444	5.2215	0.3526	0.3057
3.0257	8.6966	0.7747	0.6321
2.9452	3.7591	0.6178	0.4823
2.8845	11.9807	1.0000	1.0000
2.3248	-13.2970	-0.9979	-0.8459
2.6581	6.2765	0.8775	0.6660
1.9971	9.0063	0.7550	0.5217
1.7569	-15.0095	-1.4213	-1.3595
1.6606	3.1925	-0.1354	-0.1280
1.6838	2.4993	-0.1354	-0.1187
1.5249	-11.4573	-1.1627	-0.9881
1.3563	-8.3727	-0.5783	-0.4674
1.1760	-14.4174	-1.2999	-1.1168
0.9611	-1.9906	-0.0448	-0.0567
1.0477	-15.0372	-1.0972	-0.9807
0.8008	-17.0310	-1.5097	-1.2679
0.7519	-11.2951	-1.2239	-1.0510
1.1663	-13.6820	-1.1014	-0.9376
0.7325	-15.8784	-1.4288	-1.1594
0.6948	-11.3513	-1.0120	-0.8396
0.5093	-14.7561	-1.2902	-1.2435
0.4425	-18.8661	-1.8003	-1.8105
0.4211	-23.0733	-1.8154	-1.6179
0.4760	-20.3138	-1.7033	-1.5810

Table B.5: *Correlation values for Sim_{res} in the M&C dataset using different IC metrics.*

	Sim_{res}		
Correlation values	**RES-***IC*	**RES-***IIC*	**RES-***IC*$_{hd}$
P&S ratings	0.8308	0.8421	0.8361
RES-*IC*	-	0.9616	0.9573
RES-*IIC*	-	-	0.9928

Table B.6: *Correlation values for $Sim_{P\&S}$ in the M&C dataset using different IC metrics.*

	$Sim_{P\&S}$		
Correlation values	**P&S-***IC*	**P&S-***IIC*	**P&S-***IC*$_{hd}$
P&S ratings	0.8655	0.8843	0.8835
P&S-*IC*	-	0.9839	0.9792
P&S-*IIC*	-	-	0.9950

Table B.7: *Correlation values for $Sim_{J\&C}$ in the M&C dataset using different IC metrics.*

	$Sim_{J\&C}$		
Correlation values	**J&C-***IC*	**J&C-***IIC*	**J&C-***IC*$_{hd}$
P&S ratings	-0.8660	-0.8805	-0.8712
J&C-*IC*	-	0.9827	0.9729
J&C-*IIC*	-	-	0.9927

Table B.8 shows the results of the significance tests for Resnik's similarity for a sample size (N) of 28 and 65.

Table B.8: *Significance values for Sim_{res} in the M&C dataset using different information content metrics.*

Values of the t-statistics for Sim_{res}			
Sample size	IC **vs.** IIC	IC **vs.** IC_{hd}	IIC **vs.** IC_{hd}
28	0.3823	0.1467	0.5180
65	0.6020	0.2310	0.8157

Table B.9 shows the results of the significance tests for Pirró and Seco's similarity for a sample size (N) of 28 and 65.

Table B.9: *Significance values for $Sim_{P\&S}$ in the M&C dataset using different information content metrics.*

Values of the t-statistics for $Sim_{P\&S}$			
Sample size	IC **vs.** IIC	IC **vs.** IC_{hd}	IIC **vs.** IC_{hd}
28	1.5496	1.3632	0.0136
65	2.4403	2.1467	0.0213

Table B.10 shows the results of the significance tests for Jiang and Conrath's similarity for a sample size (N) of 28 and 65.

Table B.10: *Significance values for $Sim_{J\&C}$ in the M&C dataset using different information content metrics.*

Values of the t-statistics for $Sim_{J\&C}$			
Sample size	IC **vs.** IIC	IC **vs.** IC_{hd}	IIC **vs.** IC_{hd}
28	0.8227	0.2078	0.8592
65	1.2956	0.3272	1.3531

Table B.11 shows a descriptive analysis of the obtained results for lin's similarity Sim_{res} in the M&C dataset using different information content metrics. Those results showed although the median for the IC_{hd} metric was a little greater than IC and IIC metrics, the standard deviation and the standard error were lower than for those metrics.

Table B.11: *Descriptive analysis of Sim_{res} in the M&C dataset using different IC metrics.*

	Sim_{res}			
	P&S ratings	**RES-**IC	**RES-**IIC	**RES-**IC_{hd}
Min	0.4211	0.0000	0.0088	0.0139
Max	3.4210	12.0677	1.0000	1.0000
Average	1.7342	4.8868	0.4087	0.3635
Median	1.5928	2.7956	0.2284	0.2036
Mode	-	2.3335	0.2179	0.1906
STDEV	1.0159	3.9585	0.3360	0.3045
STDER	-	0.5763	0.5584	0.5691

Table B.12 shows a descriptive analysis of the obtained results for lin's similarity $Sim_{P\&S}$ in the M&C dataset using different information content metrics. Those results showed although the median for the IC_{hd} metric was a little greater than IC and IIC metrics, the standard deviation and the standard error were lower than for those metrics.

Table B.12: *Descriptive analysis of $Sim_{P\&S}$ in the M&C dataset using different IC metrics.*

	$Sim_{P\&S}$			
	P&S ratings	**P&S-**IC	**P&S-**IIC	**P&S-**IC_{hd}
Min	0.4211	-23.0733	-1.8154	-1.8105
Max	3.4210	12.0677	1.0000	1.0000
Average	1.7342	-4.9821	-0.4250	-0.3833
Median	1.5928	-9.8339	-0.7881	-0.6535
Mode	-	-	1.0000	1.0000
STDEV	1.0159	11.5451	1.0258	0.9222
STDER	-	0.5186	0.4708	0.4706

Table B.13 shows a descriptive analysis of the obtained results for lin's similarity $Sim_{J\&C}$ in the M&C dataset using different information content metrics. Those results showed although the median for the IC_{hd} metric was a little greater than IC and IIC metrics, the standard deviation and the standard error were lower than for those metrics.

Table B.13: *Descriptive analysis of $Sim_{J\&C}$ in the M&C dataset using different IC metrics.*

	$Sim_{J\&C}$			
	P&S ratings	**J&C-IC**	**J&C-IIC**	**J&C-IC_{hd}**
Min	0.4211	0.0000	0.0000	0.0000
Max	3.4210	23.0733	1.8669	1.8736
Average	1.7342	9.8690	0.8520	0.7729
Median	1.5928	12.6295	0.9895	0.8482
Mode	-	0.0000	0.0000	0.0000
STDEV	1.0159	7.7411	0.6846	0.6069
STDER		0.5177	0.4908	0.5092

Table B.14 shows the normalized values of the attributes used for the regression analysis of the IC_{hd} approach. The rows are presented in ascending order depending on the second last column (log-hypo(lcs)). The first two columns contain the specific synset used for each word pair of the R&G dataset. The next three columns (3 to 5) present the depths of each word and their lowest common superset respectively, normalized with max_{depth} according to Table 3.1. Columns 6, 7 and 8 show the value of the logarithm of the number of hyponyms of each word and their lowest common superset respectively, normalized with the logarithm of max_{wn} according to Table 3.1. Last column *misplaced* shows how far the semantic similarity similarity measure Sim'_{res} ranked the similarity between *word1* and *word2* from the human judgment ranking.

Table B.14: *Normalized values of the attributed used for the regression analysis of IC_{hd} approach*

word1	word2	depth(c1)	depth(c2)	depth(lcs)	log-hypo(c1)	log-hypo(c2)	log-hypo(lcs)	misplaced
midday#n#1	noon#n#1	0.63	0.57	0.69	0	0	0	2
gem#n#3	jewel#n#2	0.16	0.03	0.31	0	0	0	3
hill#n#5	mound#n#1	0.63	0.57	0.69	0	0	0	17
cock#n#4	rooster#n#1	1	1	1	0.09	0.1	0.06	1
cemetery#n#1	graveyard#n#1	0.53	0.46	0.62	0.09	0.1	0.06	4
asylum#n#2	madhouse#n#1	0.63	0.68	0.69	0.09	0	0.06	7
grin#n#1	smile#n#1	0.35	0.25	0.46	0.15	0.17	0.1	6
brother#n#5	monk#n#1	0.72	0.57	0.69	0	0.21	0.12	11
forest#n#2	woodland#n#1	0.16	0.03	0.31	0.24	0.27	0.16	1
autograph#n#2	signature#n#1	0.44	0.25	0.46	0	0.3	0.17	3
magician#n#2	wizard#n#2	0.25	0.14	0.38	0.28	0.32	0.19	2
coast#n#1	shore#n#1	0.25	0.03	0.31	0.22	0.32	0.19	3
serf#n#1	slave#n#1	0.44	0.03	0.31	0.09	0.36	0.21	3
cushion#n#3	pillow#n#1	0.35	0.35	0.46	0.33	0.17	0.21	7
glass#n#2	tumbler#n#2	0.44	0.46	0.54	0.38	0	0.25	2
journey#v#1	voyage#v#1	0	0	0.18	0.38	0.19	0.25	5
boy#n#1	lad#n#2	0.25	0.25	0.38	0.4	0	0.26	11
automobile#n#1	car#n#1	0.72	0.68	0.77	0.51	0.56	0.33	10
cord#n#1	string#n#1	0.35	0.35	0.46	0.54	0.17	0.36	2

Continued on Next Page...

Table B.14 – Continued

word1	word2	depth(c1)	depth(c2)	depth(lcs)	log-hypo(c1)	log-hypo(c2)	log-hypo(lcs)	misplaced
oracle#n#1	sage#n#1	0.44	0.25	0.31	0.19	0.17	0.45	2
coast#n#1	hill#n#1	0.25	0.14	0.23	0.22	0.32	0.46	17
mound#n#2	shore#n#1	0.35	0.03	0.23	0.19	0.32	0.46	21
implement#n#1	tool#n#1	0.35	0.35	0.46	0.9	0.89	0.59	5
bird#n#1	crane#n#5	0.63	0.89	0.69	0.92	0.1	0.6	1
bird#n#1	cock#n#4	0.63	1	0.69	0.92	0.1	0.6	1
crane#n#5	rooster#n#1	0.91	1	0.69	0.09	0.1	0.6	2
crane#n#4	implement#n#1	0.53	0.25	0.38	0.19	1	0.77	3
mound#n#1	stove#n#2	0.63	0.46	0.38	0	0	0.77	3
automobile#n#1	cushion#n#1	0.72	0.46	0.38	0.51	0.1	0.77	24
monk#n#1	oracle#n#1	0.35	0.35	0.23	0.19	0.21	0.79	2
lad#n#1	wizard#n#1	0.25	0.14	0.23	0.09	0.1	0.79	3
magician#n#1	oracle#n#1	0.35	0.35	0.23	0.15	0.21	0.79	9
brother#n#1	lad#n#1	0.44	0.14	0.23	0.19	0.1	0.79	9
sage#n#1	wizard#n#1	0.35	0.14	0.23	0.15	0.1	0.79	15
monk#n#1	slave#n#1	0.35	0.03	0.23	0.19	0.36	0.79	17
boy#n#1	sage#n#1	0.25	0.25	0.23	0.4	0.17	0.79	22
glass#n#5	magician#n#1	0.35	0.25	0.15	0	0.17	0.8	7
glass#n#2	jewel#n#1	0.44	0.46	0.31	0.38	0.35	0.83	2
furnace#n#1	implement#n#1	0.63	0.25	0.31	0.42	1	0.83	8
furnace#n#1	stove#n#1	0.63	0.78	0.31	0.42	0.32	0.83	13
asylum#n#1	fruit#n#2	0.44	0.35	0.31	0.19	0	0.83	14
cushion#n#1	jewel#n#1	0.53	0.46	0.31	0.09	0.35	0.83	15
fruit#n#2	furnace#n#1	0.44	0.57	0.31	0	0.47	0.83	20
food#n#3	fruit#n#3	0.25	0.46	0.15	0.09	0	0.84	33
boy#n#1	rooster#n#1	0.53	1	0.38	0.4	0.1	0.86	8
asylum#n#1	monk#n#1	0.44	0.57	0.23	0.19	0.21	0.91	6

Continued on Next Page...

Table B.14 – Continued

word1	word2	depth(c1)	depth(c2)	depth(lcs)	log-hypo(c1)	log-hypo(c2)	log-hypo(lcs)	misplaced
automobile#n#1	wizard#n#1	0.72	0.46	0.23	0.51	0.1	0.91	13
shore#n#1	woodland#n#1	0.16	0.03	0.15	0.28	0.27	0.92	2
coast#n#1	forest#n#2	0.25	0.03	0.15	0.22	0.27	0.92	4
graveyard#n#1	madhouse#n#1	0.53	0.68	0.15	0.09	0	0.92	5
bird#n#1	woodland#n#1	0.63	0.03	0.15	0.92	0.27	0.92	6
cemetery#n#1	woodland#n#1	0.53	0.03	0.15	0.09	0.27	0.92	6
asylum#n#1	cemetery#n#1	0.44	0.46	0.15	0.19	0.1	0.92	7
hill#n#1	woodland#n#1	0.25	0.03	0.15	0.28	0.27	0.92	9
forest#n#2	graveyard#n#1	0.16	0.46	0.15	0.24	0.1	0.92	12
cemetery#n#1	mound#n#1	0.53	0.57	0.15	0.09	0	0.92	29
food#n#1	rooster#n#1	0.16	1	0.08	1	0.1	0.94	3
noon#n#1	string#n#4	0.63	0.25	0.08	0	0	0.94	3
cord#n#2	smile#n#1	0.35	0.25	0.08	0	0.17	0.94	4
rooster#n#1	voyage#n#1	1	0.68	0	0.09	0.17	1	0
autograph#n#1	shore#n#1	0.25	0.03	0	0.27	0.32	1	3
grin#n#1	implement#n#1	0.35	0.25	0	0.15	1	1	10
grin#n#1	lad#n#1	0.35	0.14	0	0.15	0.1	1	11
car#n#1	journey#n#1	0.72	0.57	0	0.51	0.59	1	16
shore#n#1	voyage#n#1	0.16	0.68	0	0.28	0.17	1	20

Table B.15: *Ranking comparison between the human judgment and the IIC and IC_{hd} approaches.*

No.	word1	word2	Ranking-IIC	Ranking-IC_{hd}
1	cemetery#n#1	graveyard#n#1	3	4
2	automobile#n#1	car#n#1	13	10
3	midday#n#1	noon#n#1	2	2
4	gem#n#3	jewel#n#2	3	3
5	cock#n#4	rooster#n#1	1	1
6	cushion#n#3	pillow#n#1	7	7
7	coast#n#1	shore#n#1	4	3
8	boy#n#1	lad#n#2	10	11
9	forest#n#2	woodland#n#1	1	1
10	journey#v#1	voyage#v#1	6	5
11	magician#n#2	wizard#n#2	2	2
12	grin#n#1	smile#n#1	6	6
13	autograph#n#2	signature#n#1	1	3
14	asylum#n#2	madhouse#n#1	7	7
15	implement#n#1	tool#n#1	5	5
16	cord#n#1	string#n#1	3	2
17	serf#n#1	slave#n#1	3	3
18	hill#n#5	mound#n#1	17	17
19	glass#n#2	tumbler#n#2	2	2
20	furnace#n#1	stove#n#1	12	13
21	oracle#n#1	sage#n#1	2	2
22	brother#n#5	monk#n#1	12	11
23	sage#n#1	wizard#n#1	16	15
24	food#n#3	fruit#n#3	26	33
25	bird#n#1	crane#n#5	1	1
26	bird#n#1	cock#n#4	2	1
27	magician#n#1	oracle#n#1	9	9
28	brother#n#1	lad#n#1	9	9
29	crane#n#4	implement#n#1	3	3
30	crane#n#5	rooster#n#1	3	2
31	cemetery#n#1	mound#n#1	28	29
32	car#n#1	journey#n#1	17	16
33	monk#n#1	oracle#n#1	2	2
34	glass#n#2	jewel#n#1	1	2
35	furnace#n#1	implement#n#1	7	8
36	hill#n#1	woodland#n#1	15	9
37	food#n#1	rooster#n#1	1	3
38	coast#n#1	hill#n#1	17	17
39	shore#n#1	voyage#n#1	23	20
40	bird#n#1	woodland#n#1	6	6
41	forest#n#2	graveyard#n#1	14	12
42	shore#n#1	woodland#n#1	9	2
43	mound#n#2	shore#n#1	21	21

Continued on Next Page...

Table B.15 – Continued

No.	word1	word2	Rank-IIC	Rank-IC_{hd}
44	lad#n#1	wizard#n#1	2	3
45	coast#n#1	forest#n#2	8	4
46	asylum#n#1	monk#n#1	8	6
47	cemetery#n#1	woodland#n#1	8	6
48	monk#n#1	slave#n#1	17	17
49	asylum#n#1	cemetery#n#1	8	7
50	boy#n#1	rooster#n#1	6	8
51	grin#n#1	lad#n#1	13	11
52	boy#n#1	sage#n#1	22	22
53	automobile#n#1	cushion#n#1	24	24
54	cushion#n#1	jewel#n#1	13	15
55	graveyard#n#1	madhouse#n#1	4	5
56	grin#n#1	implement#n#1	8	10
57	glass#n#5	magician#n#1	12	7
58	mound#n#1	stove#n#2	15	3
59	cord#n#2	smile#n#1	2	4
60	automobile#n#1	wizard#n#1	13	13
61	autograph#n#1	shore#n#1	3	3
62	noon#n#1	string#n#4	1	3
63	fruit#n#2	furnace#n#1	23	20
64	rooster#n#1	voyage#n#1	0	0
65	asylum#n#1	fruit#n#2	19	14

Appendix C

Details of Experimental Results. The Menendez-Ichise Model

Table C.1: *Correlation values for Sim'_{length} in the M&C dataset compared with PATH.*

Correlation values	**PATH**	Sim'_{length}
P&S ratings	0.8401	0.8571
PATH	-	-0.9806

Table C.2: *Significance values for Sim'_{length} when compare with the original similarity PATH.*

Values of the t-statistics	
Sample size	**PATH vs.** Sim'_{length}
28	0.0829
65	0.1305

Table C.3: *Descriptive analysis for Sim'_{length} and the original PATH similarity.*

		Original	**Modified**
	P&S ratings	**PATH**	Sim'_{length}
Min	0.4211	0.0417	-0.5583
Max	3.4210	1.0000	0.2250
Average	1.7342	0.3504	-0.1058
Median	1.5928	0.2000	-0.0200
Mode	-	0.5000	-0.2545
STDEV	1.0159	0.3156	0.2315
STDER	-	0.5616	0.5333

Table C.4: *Correlation values for Sim'_{lch} compared with the original Sim_{lch}.*

Correlation values	Sim_{lch}	Sim'_{lch}
P&S ratings	0.8293	0.8296
Sim_{lch}	-	–0.9998

Table C.5: *Significance values for Sim'_{lch} when compare with the original similarity Sim'_{lch}.*

Values of the t-statistics	
Sample size	Sim_{lch} **vs.** Sim'_{lch}
28	0.0012
65	0.0019

Table C.6: *Descriptive analysis for Sim'_{lch} and the original Sim_{lch} similarity.*

		Original	**Modified**
	P&S ratings	Sim_{lch}	Sim'_{lch}
Min	0.4211	0.5108	-3.7658
Max	3.4210	3.6889	-0.6848
Average	1.7342	2.2331	-2.3576
Median	1.5928	2.0794	-2.2077
Mode	-	2.9957	-3.1151
STDEV	1.0159	0.9477	0.9138
STDER	-	0.5785	0.5781