DEPARTMENT OF COMPUTER LANGUAGES AND SYSTEMS

eman ta zabal zazu

FACULTY OF COMPUTER SCIENCE

FORMALIZATION OF CONCEPT-RELATEDNESS USING ONTOLOGIES: CONCEPTUAL DENSITY

applications
in the construction of lexical knowledge bases,
word sense disambiguation
and
automatic spelling correction

Eneko Agirre Bengoa

A dissertation in Computer Science

Donostia, 1998

DEPARTMENT OF COMPUTER LANGUAGES AND SYSTEMS

eman ta zabal zazu

FACULTY OF COMPUTER SCIENCE

FORMALIZATION OF CONCEPT-RELATEDNESS USING ONTOLOGIES: CONCEPTUAL DENSITY

applications
in the construction of lexical knowledge bases,
word sense disambiguation
and
automatic spelling correction

Dissertation written by **Eneko Agirre Bengoa,** under the supervision of **Kepa Sarasola Gabiola** and **Arantza Diaz de Ilarraza**.

Donostia, 1998

ABSTRACT

People decide in the most natural way up to which point two things are related or not. We call this ability measure of relatedness, that is, the measure of the strength for the relation between two (or more) words. In order to formalize and implement the measure of relatedness, structured lexical resources are needed. The main contributions of this dissertation are the following:

1. The formalization of knowledge-based relatedness among words and concepts.

2. A method to enrich and strengthen structured lexical resources extracted from dictionaries.

The first contribution yielded Conceptual Density, a measure of relatedness implemented on *WordNet,* a lexical knowledge-base for English. The theoretical advantages of our formalization are presented, as well as the evaluation results on two practical tasks. On the one hand, we have performed free-text Word Sense Disambiguation, applying Conceptual Density on all nouns appearing in a public-domain English corpus. Our results compare favorably with two other state-of-the-art techniques applied to the same corpus. On the other hand, we also tackled the automatic correction of spelling errors for English, but in this case, using Conceptual Density alongside other complementary knowledge sources and techniques, i.e. Constraint Grammar and word and co-occurrence statistics. The implemented system demonstrates that the intended correction proposal can be automatically selected with high precision.

As regards the second main contribution, it is well known that the information extracted form dictionaries has its shortcomings. The hierarchies obtained are usually hierarchies of words, not of word-senses. Moreover, the hierarchies tend to be shallow and small, with unsatisfactory structure in the higher part. The method presented in this dissertation shows that this shortcomings can be overcome on a French monolingual dictionary, *Le plus Petit Larousse.* The method comprises techniques to sense-disambiguate the genus terms in the definitions, thus producing hierarchies of word-senses, and techniques to link the senses of the entries in the target dictionary to the senses in a lexical knowledge-base in a different language, via a bilingual dictionary. Conceptual Density is the key component in both tasks. The proposed method is also used to solve the difficulties posed by the cycles in the hierarchies and the isolated entries which are unconnected. The method enables the production of high quality structured lexical resources for non-English languages, in addition to multilingual links among resources in different languages.

ACKNOWLEDGEMENTS

This thesis would not exist if it was not for the IXA research group.
The philosophy of this group and its motivation to research on Basque are highly contagious.
I specially thank the attention received by Kepa and Arantza, as well as the work done with Koldo.
To collaborate with German in the hot desert has also been decisive
in this long research-training process.

If some helped researching, others managed to keep me away, for my own benefit.

parents and sister; the *koadrila* in eibar; those in my flat who promote laziness; the computer science department of UEU; the large and tireless fun-loving group in the faculty, capable of playing *mus* after a tiring research day; the *pala*-loving club; the noisy PhD students from the overcrowded *bulego-kalte* and *fosa-común*; the potato-omelet smell filtering in from my window at midday; and more seriously, the basque-language pressure group and the insubordination to military service group (drawing posters is fun!).

after all, ana, not everything is work, is it?

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

LIST OF TABLES

LIST OF COMMON ABBREVIATIONS

AR. Association Ratio

CD. Conceptual Density

CG. Constraint Grammar

DKB. Dictionary Knowledge-Base

LKB. Lexical Knowledge-Base

LPPL. *Le Plus Petit Larousse*

MI. Mutual Information

NLP. Natural Language Processing

OFED. Oxford French/English Dictionary

WN. WordNet

WSD. Word sense Disambiguation

I.Chapter
INTRODUCTION

II.Motivation

People decide in the most natural way up to which point two things are related or not. What is more related to sheep: cow, codfish or radio? We have no problems to recognize the relations existing among the objects involved, and we are readily prepared to answer such questions. However, as it happens with most abilities connected to common sense, it is very difficult to make computers show this ability. They lack the clues to answer this kind of questions. They do not know what sheep, cow or radio are, nor do they know which are the relations among them. Were they able to answer such questions, computers could get a grasp of this area of common sense, and this could be applied to several interesting applications. We will focus on Natural Language Processing (NLP). As many people think, the key of semantic processing lies in the ability to answer such questions regarding the relatedness degree. We will call this ability relatedness, that is, the measure of the strength for the relation between two (or more) words. This measure is usually defined just for nouns.

In literature, different ways to formalize relatedness have been studied. In some works, only relatedness between words has been developed, but many others work with word senses or concepts[1]. The first ones do not distinguish between the different senses of the word *bank*, for instance. The latter, on the contrary, asked whether the words *bank* and *river* are closely related or not, would answer that "it depends": if this *bank* is riverside, then yes, but if it is a building having to do with money, then no, these are not so closely related. From our point of view, the formalization for word senses is more interesting than just the relatedness between words.

[1] Word sense and concept will be used as equivalent terms in this dissertation.

Formalizations can be also classified according to the lexical resource they use:

- Those using collections of written texts, corpora
- Those using dictionaries, specially the definition texts
- Those using structured knowledge, such as Dictionary Knowledge Bases (DKB), Lexical Knowledge Bases (LKB) and ontologies

After studying the three resource types, we deemed most reasonable to base our formalization on structured knowledge. All lexical resources keep interesting information, which is highly complementary. However, relatedness measures based on ontologies have the strongest tradition, rooted in research on psychology and artificial intelligence. The fact of having a wide coverage LKB like WordNet (Miller et al. 1993b) accessible has allowed us to apply our theoretical ideas on practice, as we implemented our measure of relatedness on this knowledge base. We will study several measures based on all different kinds of lexical resources in section III.A, and the reasons to prefer those based on ontologies will be exposed in section III.D. The measure of relatedness introduced in this dissertation is called Conceptual Density (CD), and it is based on the hierarchy of ontology concepts[2]. Even if it has been implemented on the hierarchy of WordNet, it can be applied on any lexical resource as long as it supplies concepts structured on a hierarchy.

The measure of relatedness among word senses is crucial or at least helpful for many applications, such as disambiguation of syntactic structures, word sense disambiguation, ontology building, learning of selectional restrictions, merging of ontologies, evaluation of ontologies, information retrieval, document retrieval and classification, concept clustering, automatic text correction, as well as general semantic interpretation.

Relatedness often appears closely linked to Word sense Disambiguation (WSD), and we have actually used this application to evaluate our formalization. We have used Conceptual Density to disambiguate among the senses of nouns appearing in free-running texts. This field is currently one of the most active areas in NLP, and continues to pose an open problem. The machine-translation systems built in the sixties, for instance, could not cope with word sense ambiguity, and that was one of the main reasons for their failure. As the implementations of relatedness started to use broader information sources, better results have been achieved in word sense disambiguation. Even

[2] Concepts that will be linked to the word senses in the lexicon.

if the current technology is not ready-usable in real applications, word sense disambiguation with a manageable error-rate for free-running text might be at hand in the near future.

The most extended approaches to WSD represent polisemy and homonymy using a closed list of word senses, and usually claim that knowledge-poor[3] techniques would suffice for the task. Of course, these mainstream approaches have been heavily disputed. On the one hand, there are those that do not conceive WSD separated from general NLP, as it would first require that all difficult problems in NLP were solved. They think that the (knowledge-based) advances on NLP will dilute the word sense ambiguity problem naturally. On the other hand, there are those who dispute the closed-list model, and advocate the dynamic nature of the lexicon. According to them, there is no way to state differences between word senses without first understanding processes such as metonymy and metaphor. Some skeptics go further, and claim that it is impossible to draw lines between word senses, and question the existence of word senses as discrete entities. In our opinion, these are all valuable criticisms that have to be taken into account. Ideally all these ideas should be integrated into the WSD system (and, at the same time, WSD should be tightly integrated with general LNP), but it can not be denied that, meanwhile, interesting results are being obtained using just the most simplistic approaches. More recently, it seems that the discussion has been taken to the practical side. For instance, Kilgarriff, who doubts about the existence of word senses (Kilgarriff, 1997a), has organized the Senseval[4] competition on WSD in 1998.

One of the most important goals of our research group is the development of help-systems for text comprehension and production. In this sense, we developed the commercial spelling checker/corrector for Basque "Xuxen". When finding a misspelling, spelling correction programs try to figure out the correct word, producing a list of correction proposals. It is up to the human user to choose the correct proposal. Even if this might suffice for text-processors, for other applications the program itself has to choose the correct proposal. One example of such an application is optical character recognition. It is known that when a text is scanned, optical character recognition introduces some errors (for example, the 'I' at the beginning of a word is often recognized as an 'l') and, in consequence, a post-process has to be performed in order to correct the recognition mistakes, usually involving a person that uses a spelling-corrector. In order to check whether such human post-processing can be eliminated or not, we have developed a system for automatic text-correction, which uses syntactic tools developed in our group and

[3] We use knowledge-poor as opposed to knowledge-based approaches. In other words, we can say that, currently, techniques using extensive knowledge are more popular and successful than knowledge-intensive methods.

[4] http://www.itri.bton.uk/events/senseval/

Conceptual Density. Incidentally, we have been able to test our relatedness measure in a different task.

Another important motivation of this thesis is the production of structured lexical resources. During the eighties, the NLP community, which was focusing mainly on syntactic issues, noticed the need of wide and rich lexical resources. If NLP-based applications were to deal with real texts, wide coverage lexicons were in hard need. Besides, it was found that many linguistic phenomena that had been described using complicated syntactic rules, had a lexical origin. Accordingly, the lexicon evolved from being a plain list of words to a rich structured system of words and word senses. Seeing things as they were, the research groups began to build these both rich and wide-coverage lexicons manually. The amount of information to be coded is really huge, and it requires quite a few person-years, which were available just to a handful of wide-scale projects, like for example, CYC (Lenat, 1995), EDR (Yokoi, 1995; EDR, 1993) or WordNet. As an alternative to fully-manual encoding, automatic or semi-automatic means to produce lexicons have also been considered, focusing on the information extraction from other lexical resources, namely, corpora and dictionaries.

Dictionaries have often been the starting point to extract Lexical Knowledge-Bases (LKB). From all the kinds of information extracted, from a lexical-semantic point of view, the most outstanding ones have been the hierarchies. Unfortunately, most research groups were only able to get hierarchies of words, as they were not capable of discriminating automatically the appropriate word sense involved in a certain node of the hierarchy. One important exception is the work (Bruce et al., 1992) done on the LDOCE dictionary (Procter, 1978). In this case word sense hierarchies were automatically built using the information that was coded in this specific dictionary to perform the disambiguation. Most of the extracted LKBs have been for English, as well as the ontologies and the LKBs that were built by hand. This leaves all other languages in a weak position when facing real-text NLP. There are two complementary solutions to this unwanted situation:

- To use the corpora and dictionaries that are available for each language in order to extract LKBs for that language.
- To use the knowledge already coded in English LKBs in order to create LKBs in a different language.

In other words, available lexical resources for the given language have to be used, of course, but whenever is possible and appropriate the knowledge coded in English LKBs should be translated and reused. In our opinion, our formalization of relatedness can help in both complementary approaches, as we will show in chapter VI.

The two main motivations for this dissertation, that of formalizing a measure for relatedness and that of building structured lexical resources, are interrelated. Relatedness for a given language can not be implemented if there are no structured lexical resources for that language, especially LKB and ontologies. Besides, it is difficult to devise means to automatically create structured lexical resources without the help of relatedness measures. Using the existing LKBs and ontologies for English it is possible to define the relatedness for English words and word senses. If this relatedness could be used to speed up the construction of structured resources for other languages and to link them to English resources, it would be possible then to absorb all the richness in English resources, and all these rich resources would be available for other languages. Following this direction of research, we performed two main tasks. On the one hand, we linked the entries of the French dictionary *Le Plus Petit Larousse* (LPPL, Larousse, 1980) to the entries in the English LKB WordNet, at the sense level, that is, French word sense linked to English word sense. On the other hand, both the information in the dictionary and the links built to WordNet were used to sense-disambiguate the hierarchies extracted from LPPL.

During the work that produced this dissertation, there were no wide and structured lexical resources for any languages other than English. We therefore defined relatedness for English senses, which we applied to word sense disambiguation and text-correction on English texts. Regarding the construction and enrichment of LKBs, our group had already extracted much lexical information from a French dictionary, producing a first version of a LKB. This dissertation describes the work performed on this LKB in order to enrich and reorganize its hierarchies. Nevertheless, Basque is the target of most research in our group. All methods and techniques developed in this dissertation were designed general enough to be used in the construction of structured lexical resources and in the implementation of a relatedness measure for Basque or any other language.

III.Goals

Answering the main motivations, the formalization of relatedness and the building of structured lexical resources, we set two main objectives to this thesis:

1. Theoretical: to define a measure of the relatedness among concepts and words based on knowledge.

2. Practical: to develop techniques to enrich and strengthen non-English structured lexical resources.

Both goals involve lexical resources. Regarding the theoretical goal, we will try to take advantage of existing structured lexical resources so as to model a specific inference type: the relatedness measure. The practical goal is concerned with the building of richer lexical resources, from dictionaries to LKBs.

In order to achieve these goals, we set three main tasks:

1. Design and implement Conceptual Density based on WordNet.
2. Link, at word sense level, the entries in the French dictionary *Le Plus Petit Larousse* to WordNet.
3. Disambiguate and strengthen the word sense hierarchies in the Dictionary Knowledge-Base already extracted from *Le Plus Petit Larousse*.

In order to accomplish tasks 2 and 3 it has been necessary to use Conceptual Density. We have to point out that once a strong sense-disambiguated hierarchy for French is constructed, it will be possible to apply Conceptual Density directly on this hierarchy, obtaining a relatedness measure for French. The following hypotheses is behind our approach:

> In order to disambiguate and enrich non-English LKBs, language-external knowledge is needed, which can be readily acquired via multilingual links (usually to English).

Besides tasks 2 and 3, we have applied and evaluated Conceptual Density on two other practical applications. We therefore performed two more tasks:

4. Application, tuning and evaluation of Conceptual Density: word sense disambiguation of nouns in running text.
5. Application and evaluation of Conceptual Density: automatic spelling correction.

In order to accomplish English WSD, we just used the implementation of Conceptual Density on WordNet. Besides the evident interest of WSD, we used it to evaluate our relatedness formalization. In fact, there is no agreed procedure to directly evaluate relatedness, and we deemed better to evaluate it on a practical and comparable application.

In order to be able to perform automatic text-correction, we have used different knowledge resources, apart from Conceptual Density, including, syntactical knowledge and statistical models of word and co-occurrence frequencies.

IV.Structure of the dissertation and English version availability

Main goals of dissertation	Tasks	Original Chapters		Whole chapter in English	Appen dix	Paper code
		I	Introduction	YES		
		II	Lexical resources	YES		
To define a measure of the relatedness among concepts and words based on knowledge.	Design and implementation of Conceptual Density based on WordNet.	III	Relatedness and Conceptual Density	YES		B.1 A.1 A.2 A.3
	Application, tuning and evaluation of Conceptual Density: word sense disambiguation of running text.	IV	Word Sense Disambiguation	NO	A	A.1 A.2 A.3
	Application and evaluation of Conceptual Density: automatic text-correction.	V	Automatic Spelling Correction	NO	B	B.1 B.2 B.3 B.4
To develop techniques to enrich and strengthen non-English structured lexical resources.	Linking, at word sense level, of the entries in the French dictionary *Le Plus Petit Larousse* to WordNet. Disambiguation and strengthening of the word sense hierarchies in the Dictionary Knowledge-Base already extracted from *Le Plus Petit Larousse.*	VI	Enriching the Dictionary Knowledge Base	NO	C	C.1 C.2
		VII	Conclusions	YES		

1st table: Structure and goals of the dissertation, including availability of the English version as a chapter or collection of papers.

Code	Erref	Title	Chapters
A.1	Agirre & Rigau, 1995	A proposal for Word Sense Disambiguation using Conceptual Distance	III and IV
A.2	Agirre & Rigau, 1996a	Word Sense Disambiguation using Conceptual Density	III and IV
A.3	Agirre & Rigau, 1996b	An Experiment on Word Sense Disambiguation of the Brown corpus using WordNet	III and IV
B.1	Agirre et al., 1994b	Conceptual Distance and Automatic Spelling Correction	III and V
B.2	Agirre et al., 1995	Lexical-Semantic Information and Automatic Correction of Spelling Errors	V
B.3	Agirre et al., 1998b	Towards a Single Proposal in Spelling Correction	V
B.4	Agirre et al., 1998c	Towards a Single Proposal in Spelling Correction	V
C.1	Rigau & Agirre, 1995	Disambiguating bilingual nominal entries against WordNet	VI
C.2	Rigau et al., 1997	Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation	VI

2nd table: Papers related to the dissertation.

The goals and tasks of this thesis, including their relation with the contents of each chapter, the availability of the English version of the chapters and the related papers (organized in appendices) are summarized in table 1.

The chapters without English translation are covered in the published papers which are included in the appendices[5]. There is one appendix for each chapter without English translation available. The papers have been coded using the appendix where they belong. The full list of papers related to this dissertation is shown in table 2, ordered according to the appendix.

Following this introductory chapter, **chapter II** deals with lexical resources. After noting the current importance of lexical resources in Natural Language Processing, the most influential and widely known resources are introduced. In the English version, only the corpora, dictionaries and structured resources that were actually used in this dissertation are mentioned.

In **Chapter III** (Relatedness and Conceptual Density) we study different ways of measuring the degree to which words and word senses are closely related and we introduce the most important contribution of this dissertation, Conceptual Density. Before presenting Conceptual Density, we study other formalizations of relatedness. When presenting the implementation of Conceptual Density, we introduce some parameter and variants that have to be evaluated empirically. Finally, we defend the superiority of ontology-based relatedness, and among ontology-based formalizations, the advantages of Conceptual Density. A full English version of this chapter is available.

In **chapter IV** (Word Sense Disambiguation) we evaluate Conceptual Density in a practical application, and, along the way, we adjust the parameters of Conceptual Density mentioned in the previous chapter, considering the results of this application. Even if the previous chapter defends the theoretical and practical advantages of Conceptual Density, we wanted to show that it also attains good results in practice. In Word sense Disambiguation we have to decide which of the senses for a word was intended for a given test occurrence. Almost all measures of relatedness have been applied to Word sense Disambiguation (mostly in noun disambiguation), and, furthermore, they have been sometimes designed specifically for this purpose. This chapter will start with a study of antecedents, underlining the need of different knowledge sources. Afterwards, we will explain the design of the experiments and the algorithm used to disambiguate with Conceptual Density. The experiment was set on an already disambiguated corpus, so as to automatically measure the precision of the system. From this corpus, we chose four text-sets, and we disambiguated all nouns

[5] The papers can also be accessed in http://ixa.si.ehu.es/.

in the sample (around 2.000 nouns), choosing the word senses from WordNet. A specific section is devoted to study the effects of the parameters and variants of Conceptual Density, and to choose the best values for the parameters. After evaluating the results, we will compare them to those of other methods. We have implemented two other ontology-based methods, obtaining worse results. Finally, the contributions of this chapter are outlined.

This chapter is not available in the English version, but it is fully covered in the papers (Agirre & Rigau, 1995; 1996a; 1996b). The first paper (Agirre & Rigau, 1995) presents some preliminary experiments, which were completed afterwards with the experiments presented in the second paper (Agirre & Rigau, 1996a). Finally, a slightly more extended version was published as an internal report (Agirre & Rigau, 1996b).

In **chapter V** (Automatic Spelling Correction) we have developed another practical application, that of automatically correcting spelling errors. In this chapter we introduce the implementation and the design of the system that tries to choose the correct proposal among a set of correction proposals. Firstly we present the literature on this subject. Afterwards, we introduce the results of the feasibility study on semantic and syntax-based correction. We concluded that it was absolutely necessary to include semantic knowledge, and put forward a proposal for the use of relatedness measures on the LKB built from *Le Plus Petit Larousse*. In the following section, the method for automatic correction is proposed, which is based on syntactic knowledge, semantic knowledge (provided by Conceptual Density for nouns) and corpus-based statistical techniques. Next, the design of the experiments is presented alongside the evaluation and comparison with others. Two kinds of corpora were used: one in which we introduced spelling errors artificially, and another with real spelling errors. Finally, the contributions of this chapter are summarized.

Regarding the English version, this chapter is fully available in the papers (Agirre, 93; Agirre et al., 1994b; Agirre et al., 1995; Agirre et al., 1998b; Agirre et al., 1998c). The preliminary ideas were presented in Spanish in (Agirre, 1993), specifically the feasibility-study and the preliminary proposal for using the knowledge in the French LKB. A reduced version was published in (Agirre et al., 1995). The proposal for using the relations in the LKB was further elaborated in (Agirre et al., 1994b). The design of the correction system and the actual experiments are described in (Agirre et al., 1998b; 1998c), being the latter the final version.

Chapter VI (Enriching the Dictionary Knowledge Base) tackles the other main objective of this dissertation, namely, that of building LKBs for non-English languages. First of all, lexical knowledge acquisition literature is reviewed, including multilingual resource linking, and the extraction of hierarchies from dictionaries. Hierarchies are usually extracted from dictionaries by analyzing the definitions of the word senses and detecting the hypernymy relation between the entry being defined and a distinguished term in the definition called *genus*. Special attention is paid to the problems presented by the hierarchies extracted from dictionaries. On the one hand, hierarchies are not usually sense-disambiguated. On the other hand, hierarchies tend to be shallow and isolated from each other, and to exhibit coherency problems in the top layer. Part of the problems of shallowness and isolation is caused by the cycles in the extracted hierarchies and the fact that some word senses are left out of the hierarchies (generally those defined using special relators, which do not contain a genus). Our position and proposal to overcome these problems is presented next.

In order to check whether it is possible to strengthen the construction of LKBs or not, we have studied the DKB extracted form *Le Plus Petit Larousse*. So as to make this DKB a LKB usable in NLP, we have to solve the shortcomings explained above. We propose an integrated solution method. Firstly, we studied the definitions producing cycles in the hierarchy and the definitions with specific relators, and we linked all these entries to an external LKB, WordNet (in fact, we linked all entries in LPPL). These links will enable us to integrate the mentioned problematic definitions in the overall hierarchies. Secondly, we automatically disambiguated the hierarchies, producing a word sense hierarchy. Finally, we used the LPPL-WordNet links to connect all the isolated hierarchies (including those produced by breaking the cycles and by specific relator definitions) taking WordNet as a reference. In other words, we connected the isolated hierarchies using the WordNet hierarchy. By the way, the top layer of WordNet is incorporated in the extracted hierarchy, solving the lack of coherence in hierarchies extracted from dictionaries.

In order to link the word senses of the DKB extracted from LPPL to WordNet, we used a bilingual dictionary and Conceptual Density, so that we can assign one WordNet concept (or more) to each word sense in LPPL. So as to disambiguate the hierarchy, we will use both the knowledge in the dictionary itself and the link to WordNet. We have implemented several independent techniques for disambiguation, including Conceptual Density, which were combined using a voting strategy.

This chapter is not fully covered in English. The work on cycles and the treatment of specific relators is not published yet in English. The two papers related to this chapter cover the method to link LPPL to WordNet (Rigau & Agirre, 1995) and the method to disambiguate the hierarchies extracted from LPPL (Rigau et al., 1997). The latter has been further improved as explained in

(Rigau et al., 1998) but these improvements have not been covered in the present dissertation. The results for the connection of the isolated hierarchies are unavailable in English.

In the **last chapter** we summarize the contributions made, organized by chapters, and propose further work.

V. Chapter
LEXICAL RESOURCES USED

V.A. Introduction

Lexical resources have been classified according to the following criteria:

1. Corpora
2. Dictionaries
3. Structured resources: Dictionary Knowledge Bases and Lexical Knowledge Bases
4. Structured resources: Ontologies

The order is given by the degree of elaboration. In corpora, there is only raw information about words. In dictionaries, the lexicographers include part of speech, usage codes, subject codes, definitions, examples, etc. Apart from words, we can also find word senses. Dictionary Knowledge Bases (DKB) try to make explicit the information implicit in dictionaries, especially in the definition text, and gather lexical information about words. Lexical Knowledge Bases (LKB) aim at providing all information that a system performing NLP needs about words in order to understand and produce texts. Ontologies, are conceptualizations about the world or a specific field, and try to represent all that needs to be known (entities, events, reasoning, … common sense) for a given or general application.

For the sake of this dissertation, and specially in chapter III, we will refer to ontologies on a more general sense, which includes all structured lexical resources, that is, DKBs and LKBs. The main reason is that we will be focusing on a relation (*is-a*, hypernym, superclass) that is common to all structured lexical resources. In ontologies, we find hierarchies of concepts, and in DKB and LKBs hierarchies of word senses. The relatedness measure that we will define can be equally applied to any of them. For the same reason, we will use word sense and concept in an interchangeable way.

Regarding the lexical resources that were used in this dissertation, WordNet is, without any doubt, the most important resource, as we will implement relatedness over the relations in WordNet. As for corpora, we have used SemCor to evaluate the results in the disambiguation of words (cf. chapter IV), and the Brown and Bank of English corpora for the evaluation of automatic correction (cf. chapter V). In chapter VI we will enrich the DKB which has been extracted from the *Le Plus Petit Larousse* dictionary, using also the *Oxford French/English Dictionary.* We will look at each resource closer in the next sections.

V.B. Brown and Semcor

The Brown corpora (Francis & Kucera, 1967) comprises around 1.000.000 words from the English of the United States. It has been taken from several samples of different written genres. Some of the examples of the genres are the following: *press-reportage*, *press-editorial, learned-scienc* and *humour* .

Semcor is a subset of the Brown corpus, which has been tagged with semantic information by the same team that designed WordNet (Miller et al. 1993). It includes 186 texts from the Brown corpora, and all adjectives, nouns, verbs and adverbs are tagged with the corresponding word sense from WordNet. We can see some data about this corpus in table 1. Except for a few words, all are tagged with an single sense. Both Brown and Semcor are freely available.

Quantity of words	359.732
Tagged with word sense	192.639
Tagged with multiple word senses	666

3rd table: Data on Semcor

In this corpus, the sentence "*The conductor said to Ritchie:*" is represented as follows (tagged according to WordNet version 1.4):

```
<s>
<stn>50</stn>
<wd>The</wd><tag>DT</tag>
<wd>conductor</wd><sn>[noun.person.1]</sn><tag>NN</tag>
<wd>said</wd><mwd>say</mwd><msn>[verb.communication.0]</msn><tag>VBD</tag>
<wd>to</wd><tag>TO</tag>
<wd>Ritchie</wd><df>person</df><sn>[noun.Tops.0]</sn><pn>person</pn><tag>NP</tag>
<wd>:</wd><tag>:</tag>
</s>
```

The tags are given using SGML. Word-forms are between `<wd> </wd>`, syntactic category is given between `<tag> </tag>`, and the semantic tag between `<sn> and </sn>` . For example, the word *conductor* is a noun (NN) and in this sentence, the sense it corresponds to is 1.`noun.person`, that is, the first sense of *conductor* with `person` as semantic code (we will refer to semantic codes from

WordNet in V.D). In the case of names, the semantic tag depends on the entity to which the proper noun refers to. For example, *Ritchie* is assigned person.

V.C. Bank of English

The COBUILD project led by the dictionary publishing company Collins[6], includes a corpus to monitor the development of English, which was built with the help of the University of Birmingham[7]. In 1996, the corpus had 320 million words and is currently under development. This corpus is not freely available, and permission has to be asked in order to see parts of the corpus.

V.D. WordNet

So as to implement the relatedness measure defined in chapter III, we had to choose a convenient structured lexical resource. WordNet (Miller et al. 1993b) has a very wide lexical coverage (126.520 words), best from freely available ontologies[8]. That was the main reason to choose WordNet. The other candidates were Mikrokosmos and Sensus. The first one has rich relations between concepts, but the lexicon is quite limited (they do not specify the amount of words, but it contains 4.500 concepts). The latter, was created joining semi-automatically Mikrokosmos and WordNet. It includes an interesting amount of words (90.000), but some errors were introduced in the hierarchy by the automatic joining algorithm. Unfortunately, no error-rate is given (Knight & Luk, 94). Finally, we have to point out that WordNet is very popular in NLP research (a full list of papers can be found in the WordNet web page) and that anyone can retrieve it via Internet[9].

WordNet is a LKB for English from the United States. It was designed following psycholinguistic principles. Each part of speech (nouns, verbs, adjectives and adverbs) is organized as an isolated relational system. These relational systems have synonym sets (*synset*) as conceptual units. If a word has multiple senses it will appear in several synsets, and if it has a single sense, in a single synset. For instance, *woman* has four senses, with different synonyms in each one:

```
1. woman, adult female
2. womanhood, woman
3. charwoman, char, cleaning woman, cleaning lady, woman
4. woman ((informal) a female person who plays a significant role)
```

[6] http://titania.cobuild.collins.co.uk/

[7] http://titania.cobuild.collins.co.uk/boe_info.html

[8] Most of wide coverage ontologies are not freely available. CYC and EDR are available, though considerably expensive. MindeNet is not available at all. Other ontologies are for internal use, and they are not prepared to be released (for example, NounSense).

[9] http://www.cogsci.princeton.edu/~wn

The 4th sense has no synonyms, and therefore, a gloss is included (these glosses can also be found for the rest of senses).

There are other lexical-conceptual relations between nominal synsets (see table 2). In the case of nouns, hipernymy is the most important, as it organizes the noun hierarchy. For instance, the word senses of *woman* have the following hypernyms:

```
woman, adult female                             => female, female person
womanhood, woman        => class, social class, socio-economic class
charwoman, char, cleaning woman, cleaning lady, woman     => cleaner
woman                                           => female, female person
```

The rest of the relations include meronymy and antonymy, but they are not so systematically developed. The only relation which is not between nouns is attribute, as it relates nouns and adjectives. For example, an attribute of *canary* is to be *small*. Each relation has also its inverse relation.

We can see the data regarding nouns for WordNet version 1.5 in table 4. Nouns have an average of 1,22 senses. Each synset has an average of 2,63 relations, which are basically hipernymy and hiponymy. Half the synsets have a relation of meronymy or holonymy, and the rest of relations appear scarcely.

		Amount	Per word	Per Synset
Nouns		87.671		
Synsets		60.631	1,22	
Relations	Hypernymy/hyponymy	122.246		2,01
	Meronymy/holonymy	35.067		0,58
	Antonimy	1.713		0,03
	Attribute	645		0,01
	Total	159.670		2,63

4th table: Some data of WordNet 1.5 for nouns

Nominal synsets in WordNet are structured in 26 semantic fields. These fields are listed in table 5. The sense of a noun in WordNet can be indicated directly, e.g. the third sense of *conductor*, or indirectly as referred to a certain semantic field, e.g. the first sense for *conductor* from the *noun.person* semantic field. Among semantic fields, *noun.Tops* is special, as it gathers the synsets in the upper layer of the hierarchy.

noun.Tops	noun.feeling	noun.possession
noun.act	noun.food	noun.process
noun.animal	noun.group	noun.quantity
noun.artifact	noun.location	noun.relation

noun.attribute
noun.body
noun.cognition
noun.communication
noun.event
noun.motive
noun.object
noun.person
noun.phenomenon
noun.plant
noun.shape
noun.state
noun.substance
noun.time

5th table: Semantic codes for nouns in WordNet

V.E. LPPL

The French dictionary *Le Plus Petit Larousse* (Larousse, 1980) is a monolingual dictionary. The data for this dictionary is shown in table 6. Our research team has carried out cosiderable research on this dictionary. First of all, we developed a Lexical Data-Base with all the information in the dictionary: entry, word sense number, part of speech, usage field, definition text and examples. The definitions were syntactically analyzed, and lexical-semantic relations were extracted. In the case of nouns, the extracted relations are the following: synonymy and antonymy, hipernymy, meronimy, lack-of, refer-to, derivation and several case relations. The extracted relations were used to build a DKB, structured as a semantic network.

	Total	Nouns
Entries	15.953	10.506
Defined word senses	22.899	13.740
Words in dictionary definitions (total)	97.778	66.323
Length of definitions (average)	3.27	3.82

6th table: Data for LPPL.

V.F. OFED

Oxford French-English Dictionary (OUP, 1989) is a bilingual dictionary of medium size. We only have the French-English part available in the machine-readable format. Table 7 shows the data for this dictionary. The dictionary has 13.030 entries. Each entry can have a single sense for the source word, or it can list more than one sense. We will call each of this bilingual senses subentry. For instance, the entry for the word *maintien* contains two subentries:

maintien n.m. (attitude) bearing; (conservation) maintenance

maintien 1: n.m. (attitude) bearing
maintien 2: n.m. (conservation) maintenance

The bilingual dictionary has 16.917 of such subentries for nouns. From another point of view, the dictionary contains 13.030 French words and 11.969 English words in the dictionary (see table 7).

	Amount of entries	Amount of subentries.
Total	21.322	31.502
Nouns	13.030	16.917
English nouns	11.969	–

7th table: Data for the bilingual dictionary OFED

The subentries include several fields: part of speech (mandatory, for instance, masculine noun, *n.m.*), semantic field (optional, it can be only one of 20 fields, for example, *comm.* in the example below, meaning commercial), a French clue (optional, for example, *attitude* and *conservtion* in the above examples, or *ressources* below), and last, but not least, the mandatory English translation or translations. The semantic field and the French clue help the user to choose the appropriate bilingual sense (subentry) for the French entry, in order to select the translation for the intended sense.

folie 1: n.f. madness
provision 1: n.f. supply, store
trésor 2: n.m (ressources) (comm.) finances

VI. Chapter
RELATEDNESS AND CONCEPTUAL DENSITY

Similarity plays a fundamental role in theories of knowledge and behavior. It serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations.

(Tversky, 1977)

The main object of this chapter is to define knowledge-based concept relatedness, by designing and implementing Conceptual Density, based on WordNet. First of all, we will present relatedness and review the most important literature on this subject, classified according to the used lexical resource. In the following section, we will present Conceptual Density and its predecessor, Conceptual Distance, both based on ontologies. On section C the implementation using WordNet will be put forward. Next, section D will show the more relevant features of Conceptual Density, comparing it with the other approaches to relatedness.

VI.A. Introduction and antecedents

Before going further in the object of this chapter, we want to define the terminology used in this work, so as to clarify the misunderstanding about similarity in the literature. The bases of this work are two main ideas, which are often confused: **similarity** and **relatedness**. The first one applies to two things that are similar one to the other, for example, a fork and a spoon. The second one is used to state that both things are related, for instance, a fork and a steak. Two similar things are related, indeed, but on the contrary, two related things do not have to be similar. In the literature, similarity is widely spread, but it is often used where relatedness should appear. We believe that in general we can talk about relatedness, being similarity a certain kind of relatedness. In some works about ontology **semantic distance** has been opposed to similarity and relatedness: two concepts with high similarity have a short semantic distance between them. Similarity and semantic distance are inversely related, and therefore, it is not necessary to define semantic distance. In the present dissertation, semantic distance will not be described, but **conceptual distance** will, as an specific implementation and measure of relatedness.

Many people think that relatedness is one of the keys to understand natural language. Key to the understanding or not, many applications of Natural Language Processing use implementations of relatedness: automatic correction (see chapter VIII), information retrieval, document indexing and retrieval (Sussna, 1993), clustering (Schütze 1992a; 1992b), disambiguation (e.g. of syntactic ambiguity –see for example prepositional phrase attachment ambiguity (Resnik, 1993)– or word sense disambiguation, cf. chapter VII), in the construction of ontologies (when constructing taxonomies –cf. chapter IX– , when learning selectional restrictions (Grishman & Sterling, 1994), when merging ontologies (Knight & Luck, 1994; Utiyama & Hasida, 1997), in ontology evaluation (Rada et al., 1989)), and also in semantic understanding (EDR, 1993).

That is why literature regarding relatedness is so wide; seldom it is the main subject of papers, and only from time to time is referred to. Most of the time the paper deals with an application which implicitly uses a measure of relatedness, without defining it as such.

It is not an easy task to classify the research on relatedness, not only because of the sheer quantity of it, but also because of the very different approaches used. In other words, it seems that each research group has found its own formalization of relatedness. All formulation have weak and strong features, which could mean that this field has not reached its maturity, but it is, nevertheless, understandable, if we bear in mind that each research group has studied relatedness from a different angle, depending on the target application. Although it is not the goal of this dissertation to examine all of them in depth, we will try to classify and study the best known and those which are more related to our work. We have used a general criterion to arrange them, depending on the resource used: ontology, electronic dictionary, corpus or a combination of them.

Other concepts have also been used for the classification of the works. To begin with, we will set the following difference regarding the relatedness between two words or two concepts:

1. **Paradigmatic relation**: As regards linguistics, it holds when a word can be substituted for another one in a sentence. Conceptually, given a specific ontological world, it happens when both concepts are of the same type or class. This is understood as similarity, since similar concepts tend to be classified under the same class.

2. **Syntagmatic relation:** As far as language is concerned, it holds when two words appear in the same textual context. In a pair of coordinates, we can say that if the paradigmatic relation is vertical, the syntagmatic one is horizontal (UZEI, 1982). Conceptually, even if they are concepts of different kinds, a relation exists between them. This is, in our opinion, relatedness. Depending on the textual context taken into account, we can further distinguish:
   - **Local syntagmatic relation**: collocations are one example, e.g. "good appetite", or the relation between verb and argument, as in "eat the ham". In these cases both terms are, generally speaking, close in textual context, and a direct syntactic relation is set between them.

   - **Global syntagmatic relation**: related words do not have to appear close to each other or in the same sentence. Here we find topic-related relations, e.g. the one existing between words referred to cookery, such as ham, stew, fork, kitchen, etc. We can say that the topic puts them into relation.

     In some cases, two words do not need to appear in the same textual context, but in contexts that share similar features, either syntactic or semantic. Therefore, if two words turn up in two similar texts, we can establish that there is an **indirect global syntagmatic relation** between them.

Even if this distinction may seem rather fuzzy, we will soon show that most of the studied systems fit clearly in one of these classes.

Another difference has to be set between word relatedness and concept relatedness. We are more interested in the second one, that is, in conceptual rather than linguistic relations. In order to see the relevance of concepts, Hirst (1987:5) states: "*Any practical NLU system must be able to disambiguate words with multiple meanings, and the method used to do this must necessarily work with the methods of semantic interpretation and knowledge representation used in the system*". Ontologies, Lexical Knowledge Bases (LKB) and Dictionary Knowledge Bases (DKB) are also usually organized according to concepts, as in WordNet: "*The most ambitious feature of WordNet, however, is its attempt to organize lexical information in terms of word meanings, rather than word forms*" (Miller et al., 1993b:3). There are some exceptions in DKB and LKBs, as some systems are unable to do sense disambiguation, but they admit the necessity of arranging the knowledge base according to concepts. For instance Richardson (1997:113) reports: "*In the future, this approach may be much more viable with a sense disambiguated LKB, which is work currently in progress.*"

Both word relatedness and concept relatedness are closely linked. Words, being linguistically similar, are also conceptually alike in one or more meanings, and, vice versa, words serving to name two similar concepts are also similar.

Taking into account all we have considered until now, we will study the relatedness measures in accordance with six features:

- Regarding the used resource: dictionary, ontology, corpus or a combination.
- Paradigmatic or (global/local) syntagmatic relatedness.
- Relatedness between either words or concepts.
- Evaluated on wide texts, just a few words, or not evaluated at all.
- Evaluated with nouns only, or with all parts of speech.
- Precision of the results: no results reported, medium, good, or excellent results.

As stated previously, the evaluation of relatedness is not easy. It is sometimes carried out with the help of ad hoc lists of related words elaborated by people, but more often the evaluation is done indirectly, taking into consideration the results obtained from applications such as word sense disambiguation, information retrieval or other ones. The problems of the former approach are that the lists produced by different researchers do no agree, as well as the lack of guidelines to construct such lists. Furthermore, when comparing the score produced by the system with that of the human-produced lists, only perfect matches are counted, even if the words that do not match are closely related.

We will now focus on the relatedness antecedents, paying special attention to the features named before, which are summarized in a single line after the exposition of each system.

VI.A.1. Antecedents based on ontologies

If ontology (see chapter V for our definition of ontology) is taken as basis, relatedness of two objects can be deducted from the information in the ontology. Tversky (1977), in the first axiomatization of similarity which came from the field of psychology, said: "*A new set-theoretical approach to similarity is developed in which objects are represented as collections of features, and similarity is described as a feature-matching process*". Therefore, he used a representation model based on features. Its measure was applied to different tasks, e.g. similarity of characters, of faces and of nations. In its evaluation, he compared his axiomatizing with people's intuition on similarity.

At that time in Artificial Intelligence, semantic networks were the most usual representation models, and similarity was developed mainly using *spreading activation* techniques on such networks (Quillian, 1968; Collins & Loftus, 1975). As for Collins and Loftus "*The conceptual network is organized along the lines of semantic similarity. The more properties two concepts have in common, the more links there are between the two nodes via these properties and the more closely related are the concepts*"[10]. They did not directly implement their model, but claimed that it followed the results of psycholinguistic experiments.

So as to make the *spreading activation* implementation easier, Rada et al. (1989) made quite a lot of work around the evaluation and merging of semantic networks. The measure of relatedness they present is named Semantic Distance: "*… [in spreading activation] semantic relatedness is based on an aggregate of the interconnections between the concepts. This is different from semantic distance which is equal to the minimal path length between two concepts*". Moreover, considering that there is a privileged relation structuring the semantic networks –the class-subclass or is-a relation–, instead of using all relations they claim that it is enough to apply the is-a relation: "*we hypothesize that […] is strong enough for the length of is-a paths to be used as a measure of semantic relatedness*". In their proposal for the distance formula (cf. 1st equation), distance between the concepts A and B is defined as the length of the shortest path of is-a[11] relations that links both concepts.

$$\mathrm{dist}(A,B) = \min_{p \in \mathrm{path}(A,B)} \mathrm{length}(p) \qquad (1)$$

The distance measure would be small for two closely-related concepts, and vice versa. No evaluation report was presented. This formula, in its simplicity, is quite often used, e.g. to merge different ontologies (Khnigt & Luk, 1994; Utiyama & Hasida, 1997).

[12]Ontology/paradigmatic/concepts/few/words/no results

Sussna (1993) developed further Rada´s idea applying it to the WordNet knowledge base, in order to perform word sense disambiguation on a document indexing application. The concepts of the knowledge base are word senses in this case, and, apart from subclass relations, he also proposes to use all the other relations in WordNet. Each relation will have a similarity weight (see $w_r(x,y)$ in the 2nd equation[13]) as, for example, concepts linked by a synonymy-relation are more similar than those linked by part-of relations (see also Tversky, 1977). The distance between two adjacent concepts in

[10] As we can see similarity and relatedness are confused in this work as well.

[11] As it is not necessary for this dissertation, we will not differentiate between class-subclass and is-a relations.

[12] These are the values for the features we have mentioned above, regarding the work of Rada et al.

[13] w_r in Sussna´s work is more complex than stated here, but, as he says "*the particular weights used* [w_r] *may not make that much difference*".

the semantic network (*w*(*x,y*) in the 2nd equation) is defined as the addition of the weights of all the relations between both concepts. In addition, the deeper the concepts are in the hierarchy, the shorter would be the distance (as captured by the divisor *d* in the equation).

$$w(x, y) = \sum_{r \in Wordnet-relation} \frac{w_r(x, y)}{d} \tag{2}$$

Therefore, the path having the smallest weight (cf. 3rd equation) will yield the distance between any two concepts.

$$\begin{gathered} \text{dist}(x, y) = \min_{(x, x_1, \ldots, x_n, y) \in \text{path}(x, y)} \sum_{i=0}^{n} w(x_i, x_{i+1}) \\ \text{where } x = x_0 \text{ and } y = x_{n+1} \end{gathered} \tag{3}$$

Sussna does not do a direct evaluation, but an indirect one, through the results obtained on a word sense disambiguation task.

Ontology/paradigmatic/ concepts/ wide/ nouns/good results

Mahesh et al. (1996; 1997) take the richness of the Microcosmos ontology as starting point, and argue that *spreading activation* performs word sense disambiguation in a blind way: " … *spreading activation … does not make use of available knowledge.*" When searching the paths between word senses, they affirm that the argument structure taken from the semantic analysis of the sentence should be considered. In other words, relatedness would measure the degree to which the selectional restrictions of verbs or adjectives hold for the chosen word senses. In order to compute this, they use concept-based selectional restrictions and the hierarchy of concepts.

Ontology/ paradigmatic and local syntagmatic/ concepts/proposal/ nouns-verbs/ no results

VI.A.2. Measures based on Electronic Dictionaries

There are no concepts in dictionaries, but word senses. However, these respond to conceptualizations made by lexicographers, and, in a big sense can be compared to ontology concepts. How can relatedness between those senses be measured? Unlike ontology-based works, there is no formalization based on psychology or knowledge, but only on practical approaches.

Regarding the type of relatedness, it can be said that indirect global syntagmatic relations are broadly used. In order to see whether two senses are related, their context is checked (as we are using dictionaries, the context is the definition of the sense). If they are similar, then the senses are taken to be related. The hypothesis sets that related senses will be defined with related words.

Lesk (1986) applied this hypothesis directly to word sense disambiguation: the relatedness measure of two senses is the amount of words shared by the corresponding definitions. The more words appear in both definitions, the more closely related both senses would be. As we will see below, his intuition has been fruitful, but it is also very weak, as it is subordinated to the actual words chosen when writing the definition. The evaluation is carried out through a sense-disambiguation task. The same method is put forward in (Cowie et al., 1992; Wilks et al., 1996), but in order to improve the efficiency when measuring the relatedness of a set of words, they use an optimization technique known as *simulated annealing*.

Dictionary/global syntagmatic/concepts/wide/ nouns/medium results

Veronis and Ide (1990) hold the same approach, but go further following a circular definition: the relatedness measure of two senses will be given by means of the addition of the relatedness measure of the words used in the definitions. In other words, now it is not necessary that the same words appear in the definition of both senses, it is enough if related words are used. And, when are two words related? When their senses are related. In order to see whether this hypothesis is useful or not, they built up a huge neural network using the terms appearing in dictionary definitions, adding links between the *definiendum* and the words in the definition[14], and tested it in a sense-disambiguation task (there is no systematic evaluation). The same approach was taken by Kozima and Furugori (1993), but with the object of improving efficiency, they compile the information into a vector-model (Kozima & Ito, 1995), similar to the model presented below –see also (Niwa & Nitta, 1994)–. They evaluate comparing similarity lists built up from people's intuition.

Dictionary/global syntagmatic/words/few/nouns/no results

Lesk´s method followed another development, which used vector-models based on co-occurrences in dictionary definitions. Wilks et al. (1990; 1996) collected word co-occurrences from the definitions of LDOCE. As definitions in LDOCE have been written using a reduced vocabulary (comprising 2781 words), co-occurrences are limited to those terms. As laid down by the authors, two words co-occur when they appear in the same definition. For codifying the co-occurrences of each word, they use a vector (see 4th formula). In this vector, there will be a value for each word in the reduced vocabulary (*N* in the 4th equation equals to the size of the reduced vocabulary, 2781), representing the co-occurrence strength for the word *w* and the *i*-th word in the

[14] Definitions were not lemmatized, nor analyzed.

reduced vocabulary. Six different formulas are put forward to measure the strength, all of them based on frequencies of words and co-occurrences. In the 5th equation, for example, the vector values are just the gross frequencies of the co-occurrences.

$$\vec{v}_w = \left(v_0^w, \cdots, v_N^w\right) \tag{4}$$

$$v_i^w = f_{w,z_i} \tag{5}$$

As to estimate the relatedness between two words, we can mathematically calculate the relation between the two corresponding vectors, using, for example, the cosine (see equation 6), but the authors also propose other three formulas. Wilks et al. go further on, getting the relatedness measure for word senses by creating a vector for each dictionary sense, summing up all the vectors for the words in their definition (cf. 7th equation). In this way, the mathematical measure of the relation between two vectors will yield the relatedness measure for two senses (it is enough to replace words for word senses in *w* and *z* of equation 6, using the vectors from equation 7).

$$\mathrm{sim}(w,z) = \cos(\vec{v}_w, \vec{v}_z) = \frac{\sum_{k=1}^{N} (v_k^a v_k^b)}{\sqrt{\sum_{k=1}^{N} v_k^a \sum_{k=1}^{N} v_k^b}} \tag{6}$$

$$\vec{v}_a = \sum_{w \in \mathrm{def}(a)} \vec{v}_w \tag{7}$$

This method, instead of measuring directly the overlap of words in definitions, uses the vectors for those words. The evaluation is not very thorough, as they carried out the sense-disambiguation of occurrences of *bank*.

Dictionary/global syntagmatic/concepts/few/nouns/good results

Richardson´s approach (1997) is quite alike to the ontology-based ones. In fact he builds up a semantic network from the definitions of two dictionaries (*LDOCE* and *Webster's 7th* Gove, 1969), after syntactically analyzing the definitions and extracting semantic relations. Each semantic relation has a weight based on frequencies. As the words in the definitions are not sense-disambiguated in this semantic network, it is not possible to measure the relatedness between two senses. Instead, it implements relatedness among words using paths of relations. The idea is similar to the *spreading activation* method: two words will be closely related if there are many relation-paths between them.

All relation-paths are not equally meaningful, and he is, therefore, in need of measuring the contribution of each type of relation. In order to weight each kind of relation, he uses an empirical method, which compiles 50.000 pairs of closely related words from a thesaurus and 50.000 pairs of non-related words. Sense ambiguity can cause errors in the paths joining two words (if a word in the path had different senses in each definition it appeared), so he is forced to apply very short paths, no longer than two definition words. The evaluation is made through the utilization of this thesaurus, applying held-out pairs not used for calculating the weights.

Dictionary/ paradigmatic and global syntagmatic/ words/ wide/ nouns/ good results

VI.A.3. Alternatives based on corpora

Researchers that advocate the use of corpora quote Firth (1957) often: "*you shall know a word for the company it keeps*". In other words, the features and meaning of a word will be given by the context where it appears, or, better, by the analysis of all the contexts the word appears in. On this basis, the following hypothesis has been set: Two words will be closely related if they come up in similar contexts. In order to analyze the relatedness between words, we only need to compare the contexts where they appear. Whether global or local syntagmatic relatedness is defined depends on the particular features used to model the context. If we want to study local syntagmatic relatedness, words with a direct syntactic relation will be used. In the case of global syntagmatic relatedness, wide windows are used (about ±50 words) without taking into account the order and using content words only.

So as to be able to develop corpus-based techniques that measure the relatedness between senses, the words in the corpus have to be labeled with senses, forming a training-corpus. This is one of the problems of corpus-based techniques, the need of extensive manual sense-disambiguation.

Mutual Information (MI) has been a simple and successful measure (Church & Hanks, 1990; Gale et al. 1992; 1993), which is founded on information theory. According to mutual information, if two words tend to appear always together in context, their relatedness would be stronger. On the contrary, if two terms never appear in the same context, their relatedness would be weaker. Church and Hanks use 100 word windows as context. In order to calculate the MI of words *v* and *w*, we have to consider the probabilities of each word and of both words appearing together (cf. 8th equation).

$$\mathrm{M\,I}(v,w) = \log \frac{\Pr(v,w)}{\Pr(v)\Pr(w)} \qquad (8)$$

The easiest way to estimate these probabilities –called *maximum likelihood estimate*– is to take the counts of each word (cf. *f* in equation 9) and divide it by the total quantity of words *N*.

$$\Pr(x) \cong \frac{f_x}{N} \tag{9}$$

Corpus/global syntagmatic/words[15] */few/nouns/good results/sparse data problem*[16]

MI is used in many applications, and the most quoted problem in literature is the estimation of the probabilities for rare events. All statistical techniques have also to face this problem, because a few words appear very often in texts, but most of them do it very rarely (in accordance with Zipf´s law). This problem is known as the *sparse data problem.* Which is the probability of occurrence for a word never seen in the corpus? And which is the probability of co-occurrence for two words that have turn out twice in the corpus but not together? It would be unfair to assign these two events 0 probability. The techniques brought into service to face this problem are called *smoothing* techniques.

Schütze (1992a; 1992b; 1998) found another alternative to the word co-occurrence method. He coded co-occurrences with vectors and measured the relatedness between words by means of the angle between the vectors (see Wilks´ method in the previous section). In order to be able to extend relatedness to word senses, he groups automatically the contexts of a word, by summing the vectors for all the words in the context and using clustering techniques. A human expert can then analyze the resulting clusters for each word, and assign a word sense to each cluster. According to the authors, this would be easier than tagging each occurrence of the word in the whole corpus one by one.

Corpus/global and local syntagmatic/words/wide/nouns/good results/no sparse data problems

From all the information in texts, MI and the vectors of Schütze only take into account the co-occurrences of words. There is doubtless more richness, e.g. syntactic structure. The syntactic structure can be reflected using very simple schemes, as part of speech labels appearing close to the target word, but argument structures (verb-object, noun-adjective, etc.) have also been used. Syntactic structure is usually represented as features, and therefore, the syntactic context of a word is expressed by syntactic features extracted from the corpus (that is, part-of-speech or argument structures found for the occurrences of the word). If the corpus is tagged with word senses, the

[15] We classify it as relatedness between words, because it is not straightforward to extend it for senses, as it would need hand-tagging.

[16] According to the relevance in corpus-based alternatives, we have added the sparse data problem as another feature.

relevant syntactic features for each sense can be thus collected, being directly used in sense-disambiguation (cf. chapter IV). So as to formalize relatedness, syntactic features must be used indirectly: words appearing in contexts with similar syntactic features would be closely related (in fact, we classify these relatedness measures as indirect global syntagmatic). That is what Grefenstette does (1992; 1996) when he defines the relatedness measure between words based on syntactic features. He makes an interesting evaluation, similar to that of Richardson, taking a thesaurus as a standard for measuring relatedness.

Corpus/global syntagmatic/words/few/ nouns/good results/sparse data problem

In some works a specific syntactic relations is used. For instance, many of the studies on the selectional restrictions of verbs (Grishman and Sterling, 1994; Lee, 1997) extract verb-object or verb-subject pairs form corpora, and try to find the class of nouns fitting best in each argument of the verb. By doing these, they define a measure of relatedness between verbs and nouns.

Corpus/global syntagmatic/words/few/nouns/good results/sparse data problem

VI.A.4. Combinations between ontologies, dictionaries and corpora

There are quite a few works advising the improvement of previous approaches. The reasons are various. Most important, all techniques above see the lexicon as a list without any semantic structure. Words and concepts are organized around classes, and lots of semantic features of a word are really features of the class. Therefore, why should we keep the information for each word, if most of it can be generalized as class features? Besides, the main problems of the statistical approach (**sparse data problem** and **the need of hand-disambiguation**) would be reduced if words were organized around classes. For example, in order to define the most typical object of the verb “to eat”, it is much better to use the class *eatable-object*, rather than listing *sandwich*, *ham*, *hake, apple*, etc exhaustively. Moreover, although *kiwi* does not appear in the corpus as an object of *eat*, if it is classified as an *eatable-object*, we will be able to infer the relation between *kiwi* and *eat.*

Some works try to induce classes from the corpus itself (for example, in the above mentioned Schütze´s work), but, it usually introduces a considerable degree of noise. Other works propose to use thesaurus or ontologies, in the search of intuitive and straightforward class definitions. Yarowsky (1992), for instance, in a work on sense-disambiguation takes as classes the ones given by Roget's thesaurus. In Roget's thesaurus (Kirkpatrick, 1987) each conceptual category is made of a list of related words. In order to know which are the typical contexts for each category, he collects contexts for each word in the category from the Grolier encyclopedia. Each context is made by the

100 surrounding words. He then selects from all the words in the context of the category, the most significant[17] ones, according to a statistical measure called *saliency* (see equation10).

$$\text{saliency}(w) = \log \frac{\Pr(w \mid c)}{\Pr(w)} \qquad (10)$$

In this work relatedness of words is not explicitly defined, but it is implicitly used as a method to label words with Roget´s class. Nevertheless, as in many works related to corpora, it is possible to infer the relatedness between words or senses from the measures given. Another example of this combined approach takes a similar measure trying to extract information for ontologies (Basili et al. 1997; 1995; Cucchiarelli & Velardi 1997).

Ontology+corpus/global syntagmatic/concepts/wide/nouns/very good results/no sparse data problems

Resnik (1993a; 1993b; 1995; 1997) proposes a different strategy to combine the information of ontology and corpus. As to measure the relatedness between two word senses, he first looks for their closest common ancestor in the hierarchy of the ontology (WordNet). Instead of measuring the distance to this common ancestor, he estimates the information content of the class represented by it and uses this as the relatedness measure (see formula 11, where *v* and *w* are nouns, and *c* the closest common ancestor).

$$\text{similarity}(v, w) = -\log \Pr(c) \qquad (11)$$

Class probabilities can be estimated using the frequencies in the corpus that the words belonging to the class have:

$$\Pr(c) \cong \frac{\sum_{w \in c} f_w}{N} \qquad (12)$$

[17] As for Yarowsky "*words that are likely to cooccur with the members of the category*".

This relatedness measure has been used to calculate the strength of the relation between noun senses and verbs, so as to induce selectional restrictions for verbs. It has also been used to perform word sense disambiguation of nouns, achieving good results in both tasks. Li and Abe (1995; 1996) also apply this approach in the induction of selectional restrictions and automatic clustering of nouns.

Ontology+corpus/paradigmatic/concepts/few/nouns/good results/no sparse data problems

Hearst and Schütze (1993) combine the relatedness measure for words introduced in Schütze´s previous work (1992a; 1992b) with the hierarchical information of WordNet. Their main target is to relate concepts that have no relation in WordNet's hierarchy. For instance *ball* and *referee* are closely related, but very loosely related in WordNet. They first group all synsets in WordNet in 726 categories, according to their position in the hierarchies. Then they use techniques similar to Schütze's to find relations among these groups. There is no systematic evaluation of results, and the authors themselves have confessed having obtained few relations. On the other hand they do not propose any new relatedness measure based on the so-built concept network.

Ontology+corpus/global and local syntagmatic/concepts/few/nouns/good results/no sparse data problems

Karov and Edelamnn (1996; 1998) count on dictionaries in order to collect the preliminary contexts (sentences in this case) that are related to a given word sense, avoiding in this way the need of hand-tagged data. They think there is a circularity in relatedness: words are related if they appear in similar sentences, and sentences are related if they contain related words. So as to break this circularity they take an iterative algorithm which reaches convergence and yields as result a relatedness measure for word senses and a set of sentences automatically tagged with word senses. The main advantage of their approach is the ability to train on fewer data.

Dictionary+corpus/global syntagmatic/concepts/few/nouns/good results/no sparse data problem

VI.B. Conceptual Density

In this section we introduce our proposal for the formalization of ontology-based relatedness. We want this formalization to meet the following conditions:

1. It is based on ontologies.
2. It measures relatedness among senses, making reference to ontology concepts.
3. It uses information from paradigmatic and syntagmatic relations.
4. It works for all open-class words[18].

[18] Mainly nouns, verbs and adjectives.

5. It is efficient, so as to be able to work with long texts.

The first two conditions are related, since ontologies involve relations between concepts. There, ontology should include paradigmatic and syntagmatic relations, so that we have as much information as possible when deciding the relatedness degree. The other advantage of ontologies is that there is no need to care for learning, that is, there is no need of previous hand-disambiguation. Finally, it has to be useful to work with adjectives, nouns and verbs, and efficient enough to work with real texts, not just with a few specific words.

VI.B.1. Two concepts: distance

We have taken as a starting point Rada´s work (Rada et al., 1989) and specially Sussna´s work (1993). As laid down in their research, relatedness can be formalized by means of the Conceptual Distance[19] between ontology concepts[20]. According to Sussna, there are two factors that have to be taken into account when calculating Conceptual Distance: the length of the relation-path between two concepts (the longer the path is, so is the distance) and the depth of the concepts (the deeper the concepts are, the shorter the distance is). Therefore we proposed the following formula in (Agirre et al, 1994c):

$$\begin{aligned}&\mathrm{Dist}(a,b) = \min_{p\in \mathrm{path}(a,b)} \sum_{c_i\in p} \tfrac{1}{\mathrm{depth}(c_i)}\\ &\text{where}\, a = c_0 \text{ and}\, b = c_n\end{aligned} \qquad (13)$$

Conceptual Distance between two concepts (*a* and *b* in the 13th equation) is given by the shortest path (*p*), as long as we calculate the length in a special way: for each concept in the path we will add the inverse of its depth in the hierarchy (for more information, see Agirre et al. 1994c).

VI.B.2. N concepts: density

Conceptual distance, as it stands, might be useful in many applications, but if we want to generalize distance between two concepts to distance among N concepts there is a combinatorial explosion. Using pairwise distance it is possible to measure the distance of N concepts by adding up the distance for all possible pairs (see Sussna, 1993). In order to compute the distance among eight

[19] At the beginning of this chapter, when defining relatedness, we have mentioned semantic distance. As semantic distance is not formalized and we have joined it to an ontology, we therefore call it conceptual distance.

[20] Relatedness and Conceptual Distance are opposed: the conceptual distance of two closely related concepts is close to zero, and the conceptual distance between two non-related concepts tends to ∞-.

concepts, for example, we will have to examine every pairwise combination[21] of eight, that is, twenty eight pairs. When computing the distance among all the nouns of a sentence, things get more difficult, because of word sense ambiguity. Let's assume that a given sentence has 8 words, and each word has 3 word senses. If we wanted to calculate the distance of all pairwise word sense combinations, we would have to try all word pairs (28 as before) for each possible sense combination (3^2): in total 252. Generally, if there are N words, having M senses each, we will have to measure the distance between two concepts $\binom{N}{2} \times M^2 = \frac{N \times (N-1)}{2} \times M^2$ times.

Besides, the comparison between concept sets gets difficult. Consider two sets of concepts, A and B, with a pair of concepts in each. For A and B it is possible to say that the two concepts in A are closer than the ones in B: we just have to compare the distance for each pair. If we add another concept to A, the distance among the three concepts will get bigger, and it will be impossible to compare the distance for this new A with the distance for B, because we are measuring distances among different quantities of concepts.

For this reason, we will add the following conditions to our measure:

6. The measure works for any number of concepts.
7. The measures for sets with different number of concepts are comparable.

Back to our first condition, we have to choose an ontology in order to apply the measure. Unfortunately, there are few ontologies which are nowadays both wide and free, being WordNet the only one with a good coverage vocabulary and freely accessible (see comments about this choice in section V.D). WordNet has been constructed with nearly no syntagmatic relations, having this an effect on one of the conditions, namely, that of using paradigmatic and syntagmatic relations. Therefore we have to alter conditions one and three:

1. It is based on the WordNet ontology.
3. It uses information from paradigmatic relations.

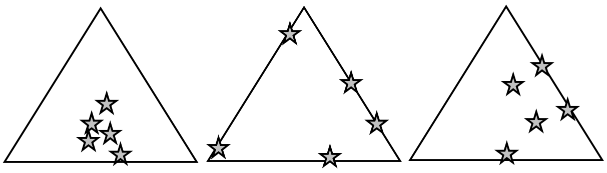
According to some authors, the fact that we stick to paradigmatic relations only is not such a hard constraint: "*we hypothesize that … is strong enough for the length of is-a paths to be used as a measure of semantic*

[21] $\binom{8}{2} = \frac{8 \times 7}{2} = 28$

relatedness" (Rada et al., 1989). The application of just hierarchy relations, in addition, has allowed us to attain a substantial improvement in efficiency, as we will see.

A measure for N concepts is not such a natural thing to develop. Up to now it was very clear that the grounds of our formulation were both the length of the path between two concepts, and the depth of the concepts in the path. However, the measure of N concepts has to look for another foundation: the abstraction to have in mind will be that of density instead of distance, that is, the amount of concepts in the subtrees of the hierarchy rather than path-length. Before going any further, we will lay some terminology. In order to differentiate them from the other concepts in the subtree, the concepts for which we are actually measuring the relatedness will be called **traces**.

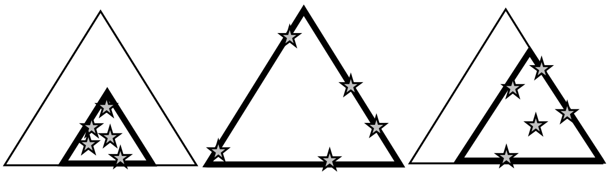
The key idea for this measure comes from the answer to this question: how many traces are needed in a subtree of the hierarchy, so as to say that the subtree is full up with traces? Or, in other words, when comparing two subtrees, how can we measure which one is fuller?



1st figure: the same subtree with three different sets of traces.

In figure 1 the same portion of the ontology appears three times, each time with a different set of traces. Would we say that the traces are equally close in the three settings? No. It seems that relatedness should be higher for the subtree on the left side, lower for the one on the middle and somewhere in between for the one on the right side. Talking about density, the highest density would be for the leftmost subtree and the lowest for the middle one. If we used Conceptual Distance of concept pairs, we would get the same result, that is, the paths between traces would be short in the leftmost subtree, and long for the traces in the middle subtree.

Leaving path lengths aside, we can observe that one of the distinguishing feature for the three sets of traces is the minimum subtree covering all five traces, as shown in the 2nd figure.



2nd figure: minimum sub-trees covering the trace sets (shown with bolder line).

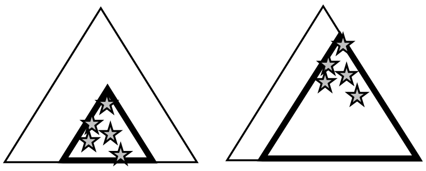
Taking in mind these sub-trees it is quite easy to see that there is a relation between relatedness among traces and the size of the minimum subtree[22]: the subtree with the highest density has the smallest size (left), and the one with lowest density the biggest size (middle). We can thus conclude that what we call density should be the relation between the amount of traces and the size of the minimum subtree covering all traces. This relation could be expressed, for example, by the amount of traces (see *a* in equation 14) divided by the size of the subtree (area(*Z*)) covering all traces (*Z* in equation 14).

$$\text{density}(Z,a) = \frac{a}{\text{area}(Z)} \tag{14}$$

Equation 14, in a first approach to density, yields the density for the subtree *Z* covering *a* traces. And, which will the density for a set of traces be? It will be given by the density of the minimum subtree covering the whole trace set *A*, or, in other words, the density of the subtree covering the trace set *A* that obtains the maximum density, as shown in equation 5[23].

$$\text{density}(A) = \max_{Z,\ \text{where}\ Z \cap A = A} \text{density}(Z, |A|) \tag{15}$$

Back to density as defined in equation 14, it takes into account the main features of Conceptual Distance: closeness and depth. The closer traces are, the smaller the area of the subclass covering the traces is (see figure 2). The same stands for depth: the deeper the traces are, the smaller the subtree is. This is shown in figure 3, where we have two sets of traces that are equally compact, but the set on the left is deeper. The set on the left will get higher density, following equations 15 and 14, because the area of the minimum subtree is smaller for the leftmost set of traces.



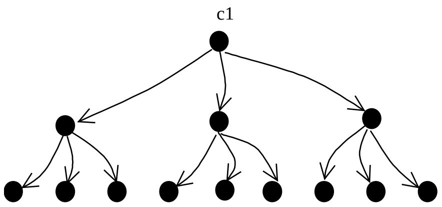
3rd figure: minimum subtrees (shown with bolder lines) covering two sets of equally compact traces.

[22] Size, area and number of nodes are equivalent ways of referring to the same measure.

[23] So as to say that subtree *Z* covers the set of traces *A*, we use *A∩Z=A. I*n order to express the amount of traces in *A*, we use its cardinal |*A*|.

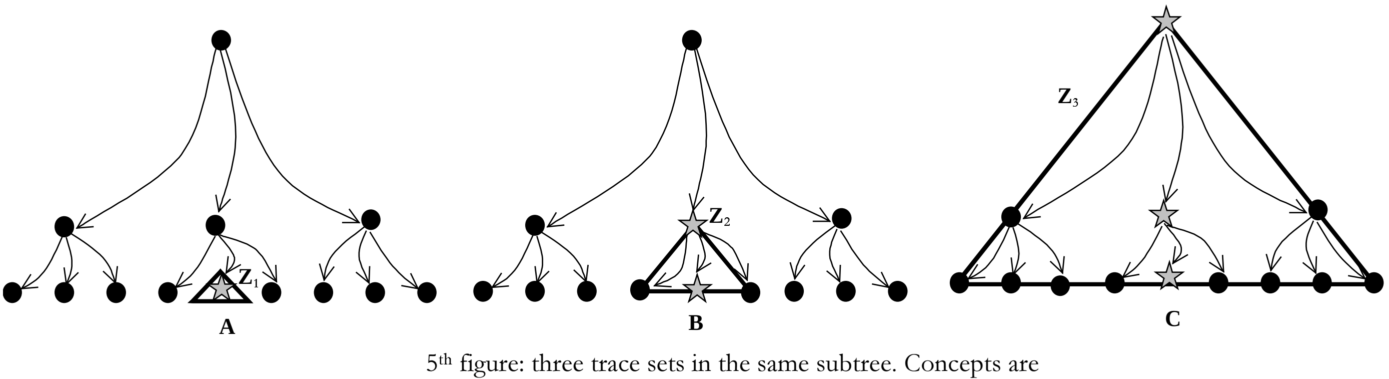
The measure defined in the 14th equation, however, has quite a lot of problems. Before analyzing them, we need to define some measures about tree topology and the relation among them: <u>the height of a subtree</u> (*h_Z*), <u>the average number of children</u> for the concepts in the subtree (μ_Z, also called branching-factor), and <u>the area of the subtree</u> (or size of the subtree, given by the amount of concepts in the subtree). The relation among these three measures –area or number of concepts, height and average number of children– is given by equation 16. An example of these measures is shown in figure 4, by means of some regular subtrees.

$$\text{area}(Z) = \text{number\_of\_concepts}(Z) = \sum_{i=0}^{h_Z - 1} (\mu_Z)^i \tag{16}$$



4th figure: the height for the subtree rooted in concept c1 (3), the average number of children of the concepts in the subtree (3), and area or number of concepts ($13=3^0+3^1+3^2$).

The problems of the 14th equation arise from the 7th condition, due to the need to compare the densities of sets with different number of concepts. Let's suppose that we want to measure the density of three different concept sets (A, B and C): one has a single trace, the other one two, and the last one three, as shown in figure 5. The subtree covering each trace is displayed with a triangle.



5th figure: three trace sets in the same subtree. Concepts are drawn as ● and traces as ☆.

According to our intuition on relatedness, would we say that the two traces in set B are more related than the three traces in set C? Or should both groups have the same relatedness? In the relatedness measure we want to formalize, we want to state that concepts from sets B and C have the same relatedness. The 14th equation, on the contrary, indicates us something different:

density(A) = density(Z_1,1) = 1/1 = 1

density(B) = density(Z_2,2) = 2/4 = 0,5

density(C) = density(Z_3,3) = 3/13 = 0,23

In our opinion, the density of all these trace sets should be 1, and to obtain this we have better not to count the traces, but to use another reference: the relation between the area and the amount of traces is not enough, and we need to consider also the height.. For instance, in figure 5, the height of subtree Z_1 is 1 and it contains one trace; the height of Z_2 is 2 and it contains 2 traces; and the height of Z_3 is 3 and it contains 3 traces; in all three the average number of children is the same.

From another point of view, what kind of weight should be given to each trace in order to make density of the trace sets in figure 5 equal to 1? Before answering this question, we will rewrite equation 14, replacing the area with the formula in equation 16, leaving a yet unknown function of the number of traces in the dividend (cf. equation 17).

$$\text{density}(Z,a) = \frac{\text{f}(a)}{\text{area}(Z)} = \frac{\text{f}(a)}{\sum_{i=0}^{h_Z - 1} (\mu_Z)^i} \tag{17}$$

Let us assume that we want to obtain the same density for all three trace sets in figure 5, and we want to make their density equal to 1. The relation we are seeking has to be established between the height and the amount of traces. As the height appears in the summatory of the divisor in equation 17, we will set f(*a*) as the formula in the dividend, but replacing the height of the tree with the number of traces *a* (cf. equation 18).

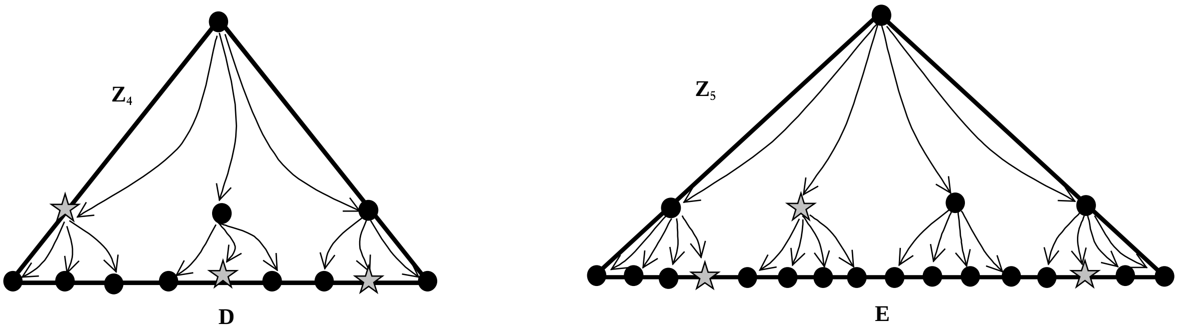
$$\text{density}(Z,a) = \frac{\sum_{i=0}^{a-1} (\mu_Z)^i}{\sum_{i=0}^{h_Z - 1} (\mu_Z)^i} = \frac{\sum_{i=0}^{a-1} (\mu_Z)^i}{\text{area}(Z)} \tag{18}$$

The divisor of the 18th equation shows the area of the subtree. The dividend shows the area that a regular subtree with the same average number of children and covering *a* traces should have in order for its density to be 1. In other words, the dividend represents a regular tree with an average number of children μ_Z that has a density of 1, and whose height equals the number of traces covered. This formula captures the relation between number of traces and area of the subtree.

The 14th equation presents yet another problem when comparing sets of traces, which is related to the topology of the subtress. As it is well known, different parts of ontologies usually have differing topologies, for example, some parts are rich in concepts, and other ones are poor. In the concept-rich parts, the average number children is bigger, being the opposite in the concept-poor parts. Let us assume that we have two concept sets located in different areas of the ontology, both with 3 traces (sets D and E in figure 6). Traces in D and E have the same distance, but, following equation 14, the densities are different:

density(D) = density(Z_4,3) = 3/13 = 0,23

density(E) = density(Z_5,3) = 3/21 = 0,14



6th figure: two different subtrees with a density of 1.

Using equation 18, however, the densities for all the trace sets considered is 1, as we wanted[24]:

density(A)=density(Z_1,1) = 1/1 = 1

density(B)=density(Z_2,2) = (1+3)/4 = 1

density(C)=density(Z_3,3) = (1+3+9)/13 = 1

density(D)=density(Z_4,3) = (1+3+9)/13 = 1

density(E)=density(Z_5,3) = (1+4+16)/21 = 1

[24] Take into account that for all subtrees, μ_Z is 3, except for Z_5, where μ_Z is 4.

In the present dissertation, therefore, Conceptual Density of concepts will be defined by means of the 15th and 18th equations.

VI.C. Implementation

Conceptual Density was implemented using the hypernymy relation in WordNet. Conceptual Distance has been implemented not only for WordNet, but also for the LPPL Dictionary Knowledge Base (Agirre et al., 1994b; 1994d). In the present dissertation we only make use of Conceptual Density, and we will not therefore define the implementation of Conceptual Distance. Before presenting the implementation we will first study some variants of Density.

VI.C.1. Variants of Conceptual Density

During the implementation we have considered that it would be interesting to study several variants and parameters of Conceptual Density. It is difficult to decide *a priori* which of the possible settings is the most convenient, and therefore, we have adopted an empirical approach, using a practical application as test-bed. The chosen application is word sense disambiguation. In this chapter the variants and parameters are introduced, but the experimental results will be shown in chapter IV (cf. Agirre & Rigau, 1996a).

VI.C.1.a) Parameter α

The formula of Conceptual Density gets in trouble when the number of traces under a subtree is too big, as the divisor in the 18th equation grows exponentially. In order to reduce this effect we added a parameter (α) to the formula, for which we found an optimal value empirically. The parameterized formula is displayed in the 19th equation.

$$\text{density}(Z,a) = \frac{\sum_{i=0}^{a-1} (\mu_Z)^{i^{\alpha}}}{\text{area}(Z)} \tag{19}$$

VI.C.1.b) How to calculate μ: μ_Z and μ_{WN}

When calculating Conceptual Density it is important to take into account the topology of the tree, which we reflect using μ_Z, the average number of children. As it can be expensive to compute μ_Z at execution time, it seems convenient to have it pre-computed and stored for each subtree in the ontology. Furthermore, we can also store the area of each subtree. When calculating density it would suffice to retrieve the area and value of μ_Z for the subtrees under consideration.

We have already studied (see equation 16) the relation among the average number of children in a subtree (μ_Z), the height of the subtree (*h_Z*) and the area (*area(Z),* number of nodes). Figure 7 shows the linear-programming algorithm in pseudocode, which given the height (H) and area (A) yields the average number of children (μ). The desired precision for the result is given as a parameter (d).

```
Input:        H height, A area
Output:       μ average number of children
Parameter:    d precision
Precondition: A>H

   if 1 <= A < H
      then μ := 1 - 1/a
      else μ := a^(1/n)
   endif
   loop
      s := μ^n;
      e := (μ*(s-A) + A - 1)/(H*s - A);
      μ := μ - e;
   until |e/μ| < d endloop
```

7th figure: algorithm for computing μ_Z

On the other hand, instead of the local measure μ_Z, we can use the global average number of children of the whole WordNet ontology (μ_{WN}). In this case we do not need to compute μ_Z for all the subtrees, but a worse measure of density is expected. In order to check whether this is the case we carried out several experiments, as reported in chapter VII (cf. Agirre & Rigau, 1996a).

VI.C.1.c) Other relations in WordNet: meronimy

Conceptual Density uses only hypernymy. However, there is another hierarchic relation among nouns in WordNet: meronymy (cf. chapter V.D). In principle, the more types of relation we consider the better results we can expect. We have empirically studied whether using meronimy improves the results or not (see chapter IV and Agirre & Rigau, 1996a). Concerning the implementation, when calculating the area of a subtree or when deciding whether a sense is covered by a subtree, meronym relations were treated as hypernym relations. The formula of Conceptual Density did not have to be changed to accommodate meronimy.

VI.C.2. Implementation on WordNet

Conceptual Density on WordNet uses just hierarchic relations, and we therefore designed an efficient algorithm that takes advantage of this.

When measuring density, we are given a set of word senses (AM), which we call traces. First of all we need to build a hierarchy, which is the subset of WordNet covering the given word senses. This subset hierarchy is built following the hypernymy links upward from the traces. All the subtrees to be taken into account will appear in this hierarchy, considering that if a subtree does not have a trace underneath, its density will be 0. Figure 8 shows the algorithm to build the hierarchy. Given a set of word senses (traces), it returns the hierarchy as defined above (H). The data structure for the hierarchy keeps a list of all the nodes in the hierarchy (H.subtrees each one representing a subtree) and, for each node, the following information: the list of direct hiponyms (H[h].hipo) and the number of traces below the node (H[h].number_of_traces), which would be used afterwards to compute density. The algorithm in figure 8 is a simplification, as it assumes that the hierarchy in WordNet follows a tree structure (a single hypernym exists for each word sense). This is not completely true in WordNet, which follows a lattice-like structure. In order to accommodate this, it is enough to change the function get_hypernymy_chain, which would return more than one hypernymy chain when the node has more than one parent.

```
FUNCTION: Build_hyerarchy(AM)
   Input:  AM set of traces
   Output: H hierarchy

      for each A in AM do
         hiper_chain := get_hypernymy_chain(A) ;
         hipo := A ;
         for each h in hiper_chain do
            push(hipo,H[h].hipo) ;
            H[h].number_of_traces ++ ;
            push(h,H.subtrees) ;
         endfor
      endfor
      return(H)
```

8th figure: building the hierarchy with the hypernyms of the traces
for which Conceptual Density has to be computed.

The implementation of the 19th equation is shown in figure 9. It computes the density of a subtree that covers a certain number of traces, given the parameter α. The arguments are the subtree itself and the number of traces underneath. It also uses the area of the subtree (Z.area) and the average number of children (Z.μ), as previously stored (see section VI.C.1.b).

```
FUNCTION:   CD(Z,A)
   Input:      Z  subtree
               A  number of traces
   Output:     CD conceptual density
   Parameter:  α
   Data:       Z.area
               Z.μ

      d1 := 0
      i := 0
      while i < A do
         d1 := d1 + Z.μ ^ (i^α)
      end
      CD := d1/Z.area
      return(CD)
```

9th figure: algorythm to calculate Conceptual Distance

Finally, in order to get to compute the Conceptual Density of a given set of traces, following the 15th equation, we will have to compute which subtree from the ones covering these traces has the highest conceptual density. The algorithm in figure 10 applies this method. It returns the density of the subtree with highest density from all the sub-trees covering all traces (`H.subtrees`).

```
FONCTION:   CD(AM)
   Input:   AM set of traces
   Output:  CD Conceptual Density

      CD := 0 ;
      H := build_hierarchy(AM) ;
      for each Z in H.subtrees do
         d := CD(Z,H[Z].number_of_traces) ;
         if d > CD then CD := d ;
      endfor
      return(CD)
```

10th figure: algorithm for the density of a set of word senses

VI.D. Evaluation and comparison with other works

Conceptual Density as defined in the present dissertation (15th and 18th equation) has not been directly evaluated. That is, we have not tested it on a list of related words to check whether our measure of relatedness and human intuition agree, following the reasons shown before (cf. section III.A for evaluation proposals). Evaluation will be carried out according to the applications where density is used, comparing our results with those obtained by other systems (cf. specially chapter VII (Agirre & Rigau, 1996a), but also chapter VIII (Agirre et al., 1998c) and chapter IX (Rigau et al., 1997)).

In this section, we will focus on the analytical comparison of the different relatedness formalizations rather than the evaluation of results. Actually, the main object will be to reason on the following argument:

> *Although the best results are not obtained in some applications, formalizations of relatedness based on ontologies are superior, both from a theoretical perspective and also because of being ready usable for different tasks. In addition, among the formalizations based on ontologies, conceptual density is more general, more efficient and the one achieving the best results.*

We will now discuss separately the two assertions, that is, the advantage of the techniques based on ontologies and the better features of Conceptual Density among ontology-based formalizations. In the following chapter we will use the results obtained on a specific application to compare Conceptual Density with other techniques.

VI.D.1. On the advantage of ontology-based techniques

As seen in the section of antecedents of this chapter, measures based on ontology have their origin in the psychology and artificial intelligence research, and these research works are the only ones studying relatedness in itself, abstracting it from specific applications.

Dictionary measures are quite *ad-hoc* in general. Corpus-based techniques are often used, but dictionary measures have an advantage over them: word senses, concepts, appear explicitly in dictionaries (for headwords generally), and the information given for a word sense can be used to characterize it. That is the foundation for the work of most of the groups: use the information about word senses given by the dictionary (definition, category, domain codes, etc.) to formalize relatedness (Lesk, 1986; Cowie et al., 1992; Wilks et al., 1996; Veronis & Ide, 1990; Kozima & Furugori; Niwa & Nitta, 1994). Karov and Edelmann (1996; 1998) do quite the same as well, but they set up a method to link the corpus and the senses in the dictionary.

We will not say that there is no information about relatedness in the dictionary, on the contrary, but this is raw information, without structure. And that is, indeed, the main contribution of the Microsoft group (Richardson, 1997). They formalize relatedness based on a Dictionary Knowledge Base constructed with relations extracted from the dictionary definitions, not directly on the raw information of the dictionary. We also set the contribution of dictionaries from this perspective, as a warehouse with the potential to produce lexical-semantic relations between word senses. We think that ontologies and dictionaries have to be joined. Word senses and concepts have to be joined, relations have to be set between sense/concepts, not between words. Chapter IX is devoted to this subject, performing word sense disambiguation on a Dictionary Knowledge Base and linking it to an external ontology.

The best results on applications using relatedness have been achieved using corpus-based measures of relatedness. Corpus-based statistical techniques are becoming very popular in the field of Natural Language Processing, and although they are mostly empiric works, a theoretical frame is also being built-up around the use of corpus. Anyway, when modeling relatedness of concepts, they have had to face several important problems. The first one is the fact that there is **not a direct definition of sense**, there is no link from words to concepts. Some works, therefore, just define relatedness between words (Grefenstette, 1992; 1996; Grishman & Sterling, 1994; Lee, 1997; Golding & Schaves, 1996). This is disturbing from a theoretical point of view, but it also brings further problems in the practical side. In order to be able to extend relatedness to word senses it demands **manual semantic tagging** of corpora (Church & Hanks, 1990; Hearst, 1991)[25]. The main trouble of manual tagging is the amount of handwork needed, as it is a time consuming task. It also rises the question of the accuracy of hand tagging, as sense boundaries are usually quite obscure, and inter-tagger agreement is usually quite low (%32 according to Jorgensen (1990)).

The improvements to the initial corpus-based proposals have been along these lines: how to avoid hand-disambiguation and how to define word sense on some more solid grounds. Schütze (1992a; 1992b) clusters automatically the contexts for a given word. A word will have as many-senses as clusters were derived. Hearst and Schütze (1993) group WordNet classes and link the occurrences of words in corpora to these classes. Accordingly, a word will have as many senses as classes to which it was linked. Yarowsky (1992) also defines word senses according to classes, but in this case with the semantic labels from Roget's thesaurus. Yarowsky himself, in later works (1994; 1995) takes another approach and presents a bootstrapping algorithm that diminishes substantially human tagging. All these works follow an interesting direction, but they never get to give a solid basis to word senses, and they fall short of linking word occurrences to ontology concepts. An attempt is presented in (Leacock et al., 1998), where WordNet is both used as dictionary, and also to diminish hand-tagging, but the results are not encouraging. Although corpus-based techniques have tried hard for many years, there are nowadays very few hand-tagged corpora, and it does not seem that corpus-based techniques will be able to go further than tagging the occurrences of a handful of words.

Corpus-based techniques also have to face the **sparse data problem**. It comes from the fact that words are taken to be isolated tokens, without considering relevant classes or sets. This, although

[25] Gale´s group (Gale et al. 1992; 1993; Yarowsky, 1993) defines senses in a different context, using the translations in parallel texts as word senses. A word will have as many senses as different translations in the parallel text. For a limited application –regarding translation– they eliminate the problem of hand-disambiguation. However, this can not be generalized to other sense or concept definitions, and the theoretical problem of defining what word senses are remains unsolved.

paradoxical in appearance, brings another problem, which can be stated as the **too-much-data problem**. On one hand, in order to alleviate the sparse data problem it is convenient to use the widest corpora possible and collect as many word occurrences as possible. On the other, all word occurrences have to be stored in order to study relatedness properly. Therefore, the information stored for each word in the lexicon is really extensive, and the information obtained for all words is huge. That is for sure, one of the reasons for evaluating corpus-based systems on small word sets[26]. Resnik (1993a; 1993b; 1995; 1997) addresses these problems using WordNet to structure word senses and words into classes. Resnik collects from corpora frequency information for the classes in WordNet, and instead of modeling word-to-word relations directly, he uses the classes (concepts) of the ontology.

As we have also said about dictionaries, we see the corpus as a huge information warehouse, but the information contained should be extracted into a structured representation. The fact that ham and fork are related can be easily derived from corpora, perhaps better than anywhere else. But saying that their relatedness weight is 0,87 should not be enough, the kind of relation should also be stated. In addition, this information should not be kept isolated, obscured in a list of co-occurrences. If the strength of the association, alongside the kind of relation was conveniently compiled, the most significant information could be incorporated into ontologies, in a more explicit and compact manner, and allowing the integration of several inference capabilities. One example of this is the above-mentioned work of Resnik, which represents the selectional restriction of verbs according to the classes of WordNet, compiling word-to-word information into classes.

Measures based on ontology, therefore, hold the strongest theoretical standpoint. Besides, word sense is clearly defined, by means of ontology concepts. The problem of ontologies, however, is one of content. Although the design of ontologies include rich relations and features, it is not easy to give values to relations and features of all concepts in the ontology. The amount of concepts should also cover a sufficient part of the lexicon. When going trough existing ontologies (cf. chapter V) we have mentioned that all ontologies have either a limited coverage of words, or just a few relations included, or both. One of the ontologies with broader lexicon is WordNet, but it mainly includes just hypernymy and synonymy relations. The problem of ontology-based relatedness measures is one of quantity of information: they can use whatever is available in their respective ontology, and no more (see proposals for further work in section X.C.1).

[26] Yarowsky (1992) for example, evaluates on the occurrences of 8 words.

VI.D.2. Conceptual Density and the other ontology-based techniques

Although the works of Tversky and Quillian are interesting, they have been placed aside when building an efficient implementation of relatedness. *Spreading activation* needs to visit all nodes and relations of the semantic network, not once, but several times.

Radar´s group, taking into account the organization of semantic networks, leaves aside all the other relations and began to use only paradigmatic relations, improving efficiency notably. Sussna was the first one to implement Conceptual Distance on WordNet, using paths between two concepts. Not only did he use paradigmatic relations, but also meronimic relations, obtaining a slight improvement in his experiments. Although the implementation has no efficiency problems when computing the distance between two concepts (it has to explore the average depth of the hierarchy twice, which can be achieved using an algorithm with constant order, O(ct)), in order to compute the distance among N words having M average senses pairwise distance has to be computed $\frac{N \times (N-1)}{2} \times M^2$ times (cf. section VI.B.2). This demands an algorithm with $O(N^2)$ complexity. Having in mind that some authors use windows with 100 words, for instance in word sense disambiguation, this problem becomes crucial.

Conceptual Density, on the other hand, computes the density for all the words under consideration at once, processing the $N \times M$ senses only once, and therefore, allowing for an algorithm with lower complexity.

As already mentioned in section VI.B.2, the problem of using pairwise relatedness is not only one of efficiency. In theoretical grounds, it is not very clear what does it mean to add pairwise distances for N concepts, which makes altogether difficult to compare distances among sets with different number of concepts. Conceptual Density, on the contrary, gives us a measure allowing to compare naturally the relatedness of concept sets with differing cardinality.

So as to finish with the examination of Conceptual Density (the evaluation related to the practical results will be given in chapters VII (Agirre & Rigau, 1996a), VIII (Agirre et al., 1998c) and IX (Rigau et al., 1997)), we will reconsider the conditions set beforehand on the goal relatedness measure:

1. It is based on ontologies.
2. Measures relatedness among senses, making reference to ontology concepts.
3. Uses information from paradigmatic and syntagmatic relations.

4. Works for all open-class words.
5. Efficient, so as to be able to work with long texts.
6. The measure works for any number of concepts
7. The measures for sets with different number of concepts are comparable with each other.

From these required features we already saw that Conceptual Density meets 1, 2, 5, 6 and 7. Regarding the 4th condition, we have only tried the Conceptual Density with nouns (cf. chapters VII, VIII and IX), but a priori there is no problem to extend it to the other parts of speech.

Regarding the 3rd requirement, it was already mentioned in chapter V that there is nowadays no freely accessible wide-coverage ontology except WordNet. Conceptual Density, therefore, has been designed having WordNet in mind, and it does only use hypernym and meronym relations. In other words, it does not use any syntagmatic relation.

VII. Chapter
WORD SENSE DISAMBIGUATION

In this chapter we evaluate Conceptual Density in a practical application, and, along the way, adjust the parameters of Conceptual Density mentioned in the previous chapter, considering the results of this application. Even if the previous chapter reasons the theoretical and practical advantages of Conceptual Density, we wanted to show that it also attains good results in practice. In Word sense Disambiguation we have to decide which of the senses for a word was intended for a given test occurrence. Almost all measures of relatedness have been applied to Word sense Disambiguation (mostly in noun disambiguation), and, furthermore, they have been sometimes designed specifically for this purpose. This chapter will start with a study of antecedents, underlining the need of different knowledge sources. Afterwards, we will explain the design of the experiments and the algorithm used to disambiguate with Conceptual Density. The experiment was set on an already disambiguated corpus, so as to automatically measure the precision of the system. From this corpus, we chose four text-sets, and we disambiguated all nouns in the sample (around 2.000 nouns), choosing the word senses from WordNet. A specific section is devoted to study the effects of the parameters and variants of Conceptual Density, and to choose the best values for the parameters. After evaluating the results, we will compare them to those of other methods. We have implemented two other ontology-based methods, obtaining worse results. Finally, the contributions of this chapter are outlined.

This chapter is not available in the English version, but it is fully covered in the papers (Agirre & Rigau, 1995; 1996a; 1996b), that can be found in appendix **A**. The first paper (Agirre & Rigau, 1995; **A.1**) presents some preliminary experiments, which were completed afterwards with the experiments presented in the second paper (Agirre & Rigau, 1996a; **A.2**). Finally, a slightly more extended version was published as an internal report (Agirre & Rigau, 1996b; **A.3**).

VIII. Chapter
AUTOMATIC SPELLING CORRECTION

In this chapter we have developed another practical application, that of automatically correcting spelling errors. In this chapter we introduce the implementation and the design of the system that tries to choose the correct proposal among the set of correction proposals. Firstly we present the literature on this subject. Afterwards, we introduce the results of the feasibility study on semantic and syntax-based correction. We concluded that it was absolutely necessary to include semantic knowledge, and put forward a proposal for the use of relatedness measures on the LKB built from *Le Plus Petit Larousse*. In the following section, the method for automatic correction is proposed, which is based on syntactic knowledge, semantic knowledge (provided by Conceptual Density for nouns) and corpus-based statistical techniques. Next, the design of the experiments is presented alongside the evaluation and comparison with others. Two kinds of corpora were used: one where we introduced spelling errors artificially, and another with real spelling errors. Finally, the contributions of this chapter are summarized.

Regarding the English version, this chapter is fully available in the papers (Agirre, 93; Agirre et al., 1994b; Agirre et al., 1995; Agirre et al., 1998b; Agirre et al., 1998c) that can be found in appendix **B**. The preliminary ideas were presented in Spanish in (Agirre, 1993)[27], specifically the feasibility-study and the preliminary proposal for using the knowledge in the French LKB. A reduced version was published in (Agirre et al., 1995; **B.1**). The proposal for using the relations in the LKB was further elaborated in (Agirre et al., 1994b; **B.2**). The design of the correction system and the actual

[27] This paper is not available.

experiments are described in (Agirre et al., 1998b; 1998c; **B.3** and **B.4**), being the latter the final version.

IX. Chapter
ENRICHING THE DICTIONARY KNOWLEDGE-BASE

This chapter tackles the other main objective of this dissertation, namely, the building of LKB for non-English languages. First of all, lexical knowledge acquisition literature is reviewed, including multilingual resource linking, and the extraction of hierarchies from dictionaries. Hierarchies are usually extracted from dictionaries by analyzing the definitions of the word senses and detecting the hypernymy relation between the entry being defined and a distinguished term in the definition called the *genus*. Special attention is paid to the problems presented by the hierarchies extracted from dictionaries. On the one hand, hierarchies are not usually sense disambiguated. On the other hand, hierarchies tend to be shallow and isolated from each other, to exhibit coherency problems in the top layer. Part of the problems of shallowness and isolation is caused by the cycles in the extracted hierarchies and the fact that some word senses are left out of the hierarchies (generally those defined using specific relators, which do not contain a genus). Our position and proposal to overcome these problems is presented next.

In order to check whether it is possible to strengthen the construction of LKBs or not, we have studied the DKB extracted form *Le Plus Petit Larousse*. As to make this DKB a LKB usable in NLP, we have to solve the shortcomings explained above. We propose an integrated solution method. Firstly, we studied the definitions producing cycles in the hierarchy and the relator type of definitions, and we linked all these entries to an external LKB, WordNet (in fact, we linked all entries in LPPL). These links will enable us to integrate the mentioned problematic definitions in the overall hierarchies. Secondly, we automatically disambiguated the hierarchies, producing a word sense hierarchy. Finally, we have used the LPPL-WordNet links to connect all the isolated hierarchies (including those produced by breaking the cycles and by specific relator definitions) taking WordNet as a reference. In other words, we connected the isolated hierarchies using the WordNet hierarchy. By the way, the top layer of WordNet is incorporated in the extracted hierarchy, solving the lack of coherence that hierarchies extracted from dictionaries exhibit.

In order to link the word senses of the DKB extracted from LPPL to WordNet, we used a bilingual dictionary and Conceptual Density, so that we can assign one WordNet concept (or more) to each sense of LPPL. So as to disambiguate the hierarchy, we will use both the knowledge in the dictionary itself and the link to WordNet. We have implemented several independent techniques for disambiguation, including Conceptual Density, which were combined using a voting strategy.

This chapter is not fully covered in English. The work on cycles and the treatment of specific relators is yet unpublished in English. The two papers related to this chapter cover the method to link LPPL to WordNet (Rigau & Agirre, 1995; **C.1**) and the method to disambiguate the hierarchies extracted from LPPL (Rigau et al., 1997; **C.2**). Both papers are included in appendix **C**. The latter has been further improved as explained in (Rigau et al., 1998) but these improvements have not been covered in the present dissertation. The results for the connection of the isolated hierarchies are unavailable in English.

X. Chapter
CONCLUSIONS

X.A. Summary

The main contributions of this work are two:

a. A formalization of relatedness: Conceptual Density

b. A method to enrich and strengthen hierarchies extracted from dictionaries

We formalized a measure for the relatedness between word senses: Conceptual Density. This measure is based on ontologies, and therefore, re-utilizes information used for general NLP. It can be applied to any ontology, it does not need any previous preparation, and it is able to operate in all the fields covered by the ontology. We implemented Conceptual Density for nouns on WordNet.

We claim that our formalization is more interesting than both measures based on other lexical resources (corpora or dictionaries) and other measures based on ontologies. We reasoned this position in chapter III, but we also tried to show its advantages in practice:

- In Word sense Disambiguation (chapter IV)
- In Automatic Spelling Correction (chapter V)

Conceptual Density performs well in word sense disambiguation of nouns, although the comparison with other systems is difficult. In order to compare them better, we implemented two other ontology-based systems, which did not perform as well as Conceptual Density on the same test-set. The results in automatic spelling correction were not so conclusive. As the current implementation of Conceptual Density only works for nouns, we could only apply it when all the correction proposals were nouns, and therefore, it was seldom used in the test corpora. The automatic spelling correction system introduced in this dissertation uses additional knowledge sources.

Concerning the second main contribution, we presented a method to enrich and strengthen the hierarchies extracted from dictionaries. This method uses both Conceptual Density on WordNet and the knowledge contained in the dictionary under study. We have improved the hierarchies extracted from the *Le Plus Petit Larousse* French dictionary in two main areas:

- Linking the entries and genus from the French dictionary *Le Plus Petit Larousse* to WordNet synsets using a bilingual dictionary.
- Sense-disambiguating and strengthening the hierarchies of the DKB extracted from *Le Plus Petit Larousse*.

Thanks to the first one, we can overcome some shortcomings of the extracted hierarchies, using the hierarchy of WordNet as a top ontology to do the following: join the definitions with specific-relators, erase the cycles in the hierarchy, join isolated mini-hierarchies and give all hierarchies a coherent top level. It also supports the disambiguation of word-based hierarchies into word sense based hierarchies. The method can be applied to disambiguate and strengthen hierarchies taken from any dictionary.

Besides, the method can be also used to join lexical resources, and it could be also used to link heterogeneous resources, in the same language or in different ones: ontologies to LKBs, LKBs to LKBs and so on. This opens new perspectives for the enriching of lexical resources, as languages poor in linguistic knowledge can absorb the knowledge built for English, provided this knowledge can be readily applied to the other language, of course. Word sense disambiguation, from this perspective, can also be cast as a method to join lexical resources, that is, to link the occurrences of the words in the corpus to word senses/concepts in the ontology. This point of view offers new paths to enrich ontologies.

Regarding future-work, we see a great demand of both wide-coverage and relation-rich ontologies. In fact, Conceptual Density as implemented in this dissertation, only takes advantage of the information in WordNet, that is, of mostly paradigmatic relations. Although we obtained good results in the tasks where Conceptual Density was applied, it is also clear that syntagmatic relations offer good perspectives of improvement, for example in word sense disambiguation, but specially in automatic spelling-correction, in order to extend the contribution of Conceptual Density.

We think that a close coordination between corpora, dictionaries and ontologies is needed to perform word sense disambiguation, but also to offer a robust solution for other lexical-semantic problems in NLP. Chapter VI (Rigau & Agirre, 1995) shows a method to join a LKB (WordNet) and a DKB (*Le Plus Petit Larousse*) in different languages. This integration can be used to enrich WordNet with the information in other LKBs or DKBs, but it would not be sufficient to gather all the needed knowledge for WSD. For instance, regarding syntagmatic relations, there are no wide-coverage lists of selectional restrictions. In oder to be able to favour their learning, we have to support the analysis and use of the definitions in the dictionaries (for example, using the techniques mentioned in chapter VI), which can be integrated in the ontologies once the words in the definitions are disambiguated. Corpora are also a valuable source of information. In chapter III we present several statistical measures based on corpora that capture quite well relatedness for words, and argues that the underlying relations should be coded in ontologies. By means of word sense disambiguation it would be possible to convert these relations between words in relations between word senses taken from a given reference ontology and, therefore, the relations could be added to the ontology. Extending Conceptual Density in an appropriate way, we would take advantage of the relations of these new ontologies, coded in a robust and efficient representation, so as to calculate relatedness using knowledge which was gathered from many different sources.

First, let's study, in more depth, the main contributions made in each chapter. Then, we will present the future-work related to each subject of this dissertation.

X.B. Contributions

X.B.1. A measure of relatedness: Conceptual Density (chapter III)

We have designed and implemented Conceptual Density, which formalizes relatedness among word senses based on ontologies. Conceptual Density takes advantage of paradigmatic relations –hypernimy and meronimy– , and works with nouns at present, although it can also be adequate for verbs.

It shares many features with other formalizations based on ontologies. Being ontologies the main model for knowledge representation in psycholinguistics and artificial intelligence[28], they have a strong theoretical basis. They offer a measure between word senses, with a solid foundation for sense differentiation, given by the senses being linked to ontology concepts. Besides, they do not require any kind of hand disambiguation, and do not show sparse-data or too-much-data problems. These are positive features as compared with other corpora or dictionary-based techniques.

[28] We want to underline that we adopt general definition of ontology, which includes all symbolic knowledge bases.

Measures based on ontologies, however, do have efficiency problems. Furthermore, the measure of relatedness is limited to two concepts, and no more. Conceptual Density overcomes these two limits. It can measure the relatedness for any number of concepts, offering the possibility of comparing the relatedness of sets with different numbers of concepts. It is efficient enough to work with large noun sets from real texts.

X.B.2. Application of CD: Word Sense Disambiguation (chapter IV)

We implemented and tested a disambiguator based on Conceptual Density, which uses the paradigmatic knowledge in WordNet. Thanks to the features of Conceptual Density, we developed a system that disambiguates according to the word senses in the ontology, and is capable of disambiguating the nouns in real running texts. It can be applied to any kind of text, without any adaptation.

As for the results of the experiment, we have proved that Conceptual Density is useful for WSD, and we have seen that it attains better results than two other formalizations of relatedness based on paradigmatic knowledge in WordNet –Sussna (1993) and Yarowsky (1992)–.

When comparing it with other experiments in the WSD literature, ours tackles the most difficult aspects of the problem: fine-grained sense distinction, real texts from different genres, all nouns in the text, leaving aside partial results and accepting one single sense. The texts chosen at random (10.000 words overall) were not at all easy to disambiguate. However, when disambiguating sense distinctions on the fine-grained level in WordNet, we obtained a precision of 64%, and one of 71% if we disambiguated to a coarser file level. Coverage is very wide, as we disambiguated 86% of the nouns in the text.

X.B.3. Application of CD: Automatic Spelling Correction (chapter V)

We designed and implemented a system that performs the automatic correction of running texts, choosing the correct proposal for non-word spelling errors. On the one hand, we proved that automatic spelling-correction is close to be a feasible task with current technologies, and, on the other, we saw that the contribution of Conceptual Density was modest.

This system combines different kinds of knowledge: syntactic (Constraint Grammar), lexical-semantic (Conceptual Density), frequency of words, context-based statistical measures and specific heuristics. Thanks to Constraint Grammar, frequency of words in documents and context-based statistics, the system is able to choose a single proposal for 24 out of 25 errors (two proposals for the rest) with 90% precision, and 100% coverage. These results prove that automatic spelling-correction can be performed nowadays with current technology.

Conceptual Density could be applied to 8% of all errors, as it is only applied whenever all proposals are nouns. Although the sample is too small to provide reliable data, it attained 75% precision. The reason for this modest performance is not CD itself, but the shortcomings of the knowledge in WordNet, as pointed down in chapter III.

X.B.4. Techniques to enrich and strengthen structured lexical resources (chapter VI)

The problems exhibited by hierarchies extracted from dictionaries are mentioned at the beginning of chapter VI, and the hierarchies extracted from *Le Plus Petit Larousse* (Artola, 1993) are not an exception. So as to solve these problems we saw the need of an external ontology, which would organize the top-levels of the hierarchies and would link the different hierarchies in a single structure. Besides, we also used the links to the external ontology in order to solve cycles in the hierarchy, and to integrate the definitions with specific-relators in the hierarchies. This external ontology has also been the key to disambiguate the words in hierarchies. We organized the overall method to strengthen and enrich hierarchies extracted from dictionaries in four parts:

X.B.4.a) Treatment of cycles and definitions with specific-relators

We introduced a method to break the cycles and to integrate them in the hierarchies, which uses the LPPL-WordNet link. Thanks to the method presented we were able to break all the cycles in the LPPL-derived hierarchies. The method to integrate the otherwise isolated definitions with specific-relators was able to link 78% of such definitions to a sense-disambiguated hypernym in the hierarchy, and 63% to a WordNet synset. The attained precision of both types of links is around 90%. As a result, all the cycles are normally integrated in hierarchies, and almost all the specific-relator definitions are either integrated in the hierarchy or linked to WordNet. Afterwards, the method for linking isolated hierarchies, will also integrate those specific-relator definitions which were only linked to WordNet.

X.B.4.b) Linking resources in different languages at a concept-level

First of all, we automatically linked the senses of a French-English bilingual dictionary to concepts of WordNet (bilingual-WordNet link), using just Conceptual Density. By means of this method, we linked 43% of the noun senses with a precision of 95%. This type of links is very important to link words from foreign languages to a given ontology. In fact, simpler methods have been used with the same goal, e.g. to join Spanish words to the Sensus ontology (Okumura & Hovy, 1994), and also, within the EuroWordNet project, to build the Spanish WordNet (Rigau & Agirre, 1995;

Atserias et al. 1997). We think that the method presented here using Conceptual Density would help to improve the precision reached in those works.

Regarding the method to join the entries and genus of LPPL to WordNet concepts (LPPL-WordNet link), the bilingual-Wordnet links have been valuable to improve the results. Apart from these links, we made use of Conceptual Density, hypernimy relations, some simple heuristics and saliency-based statistical information, including also the treatment of the specific-relator kind of definitions. Altogether, we have been able to link 87% of the noun senses of the entries in LPPL to WordNet synsets, with a precision of 80%. Both Conceptual Density and the heuristic using hypernimy relations are based on the paradigmatic links of WordNet. The technique using saliency employs statistical measures on the words in the definitions and the semantic codes in WordNet.

X.B.4.c) Genus disambiguation

In this work, we have shown that genus disambiguation is not only limited to special English dictionaries such as LDOCE, since we developped a method that attains a precision of 82% on the hierarchies of LPPL. This method can be applied to any other dictionary, as the results obtained with a Spanish dictionary – 83% precision– show (Rigau et al. 1997).

X.B.4.d) Linking isolated hierarchies extracted from dictionaries

Hierarchies derived from dictionary definitions, even after disambiguation, exhibit several deficiencies: most of them are small and isolated from each other, without any link between them. Besides, it is also known that the top layer of such hierarchies is not very adequate. We have proposed a method that tries to solve both problems, taking advantage of the links to WordNet that were already computed. In this procedure we link the root of the isolated hierarchies to WordNet, using the upper layer of WordNet to provide a coherent upper level to our hierarchy as well, and by the way linking all isolated hierarchies to each other via WordNet relations. The proposed method is general, and it will be also possible to join the hierarchies extracted from dictionaries to any ontology, giving us the opportunity of choosing the most interesting top level.

X.C. Future Work

X.C.1. Improvement of Conceptual Density (chapter III)

We see three main avenues to improve Conceptual Density:

- Regarding the information used: To either obtain a richer ontology providing syntagmatic relations and selectional restrictions, or to enrich WordNet with those relations from elsewhere. Unfortunately, this information is not readily available at present, but methods to extract them automatically from dictionaries and corpora are being studied. We have already mentioned in section X.B.4, for instance, that it is possible to extract syntagmatic relations from the analysis of the differentia in dictionary definitions. In the chapter on automatic spelling correction (chapter V), we have also seen that the raw information gathered by context statistics from corpora hides implicit syntagmatic relations and selectional-restrictions. Thanks to the integration of lexical resources (chapter VI) and word sense disambiguation (chapter IV), it would be possible to integrate this information in the knowledge-base of WordNet.

- Regarding the formula: To change the Conceptual Density formula, so that it includes syntagmatic relations. In section V.B.2, we have shortly described how syntagmatic relations can be integrated in Conceptual Density, along the lines proposed in (Agirre et al. 1994b) for an efficient conceptual distance concerning both paradigmatic and syntagmatic relations from LPPL.

- Faster implementation: Even if the complexity of the Conceptual Density algorithm is acceptable, we think that a faster implementation can be obtained. One of the reasons for that is that LISP has been the implementation language, and the other one, that the access to the information in WordNet is not optimized. The research group of the Electricity and Electronic Department of the UNED is developing a version on C++, within the ITEM[29] project. This version will be soon integrated in the GATE[30] environment for linguistic engineering (Cunningham et al. 1997), in the module for word sense-disambiguation.

X.C.2. Word Sense Disambiguation (Chapter IV)

The design of the experiments could be improved as follows:

- Disambiguating text chunks in one go, following discourse-structure. This way, instead of disambiguating words one by one using a context window, whole parts of the text, e.g. paragraphs, can be disambiguated altogether, improving efficiency. Besides, precision would also improve, as unrelated text parts would be treated separately.

- It would be interesting to study whether there is any correlation between the measure of density and the correct choice of sense. If that was the case, we would leave the cases with density below a certain measure ambiguous, and precision would improve (at the cost of a lower coverage).

[29] http://sensei.ieec.uned.es/item/

If we want to build a more powerful system for WSD, in addition to the improvements to Conceptual Density outlined in the previous section, it is necessary to supplement relatedness measures with other useful information sources: For instance, frequencies of word sense, both overall and local to the text we are disambiguating, whether the sense appears always as a collocation, information about the syntactic structure around the word sense, and so on. We would thus build a more thorough system for sense-disambiguation, which would code lexical-semantic information by means of Conceptual Density, and which would be able to combine this with other knowledge.

While this dissertation is being written, we are also preparing the SENSEVAL competition[31]. Many groups world-wide are going to present their systems. For this competition, we will try to combine Conceptual Density with several dictionary techniques (related to those used in chapter VI) and present a disambiguation system that does not need any training. We also plan to present an additional system, which will combine the previous with a context-based trainable system (cf. chapter V).

X.C.3. Automatic Spelling Correction (Chapter V)

When designing the experiment we did not bear in mind that the learning corpus (Brown) and the testing corpus (Bank of English) were from different dialects. It is for sure that this mismatch affects negatively to the results of the overall frequency and techniques based on context-statistics. The best solution would be to learn from the held-out data of the Bank of English, but, unfortunately, there are serious limitations to get the data. Consequently, the corpus of real-errors had a very small context window around the error (more or less one sentence). This has seriously damaged the heuristic that proved to be most powerful, i.e. the document frequency, which needs to gather frequencies from whole documents, not just the sentence around the error. We are trying to overcome these limitations, which would improve strongly our results.

In order to improve precision we should refine the knowledge used. Constraint Grammar, for example, can be better adapted to deal texts with misspellings, since the version we used was not designed for that. Conceptual Density, would also get better results, specially in coverage, if WordNet was enriched with syntagmatic relations, allowing to tackle proposals from different categories.

[30] http://www.dcs.shef.ac.uk/research/groups/nlp/gate/

[31] http://www.itri.bton.ac.uk/events/senseval/cfp2.html

Finally, the results in this task do not ratify (nor deny) one of the features of Conceptual Density that we mentioned in chapter III, i.e. the fact that it can also be used to measure relatedness between words. In the algorithm for automatic spelling correction we have chosen the proposal that had the word sense with the highest density, but we should also try other possibilities like, for instance, adding the densities for all word senses of each proposal, and choosing the proposal with the highest overall density.

X.C.4. Strengthen and enrich lexical resources further (Chapter VI.)

X.C.4.a) Multilingual links between concepts

Using wider bilingual dictionaries would improve the coverage and precision in the LPPL-WordNet link. On the one hand, we would have a wider Bilingual-WordNet link (enabling for more coverage and precision in the LPPL-WordNet link). On the other, the lack of translation for a word sense in LPPL is a serious error-source, and a wider bilingual dictionary would reduce those (better precision).

Another opportunity to raise the coverage of the Bilingual-WordNet link is given by the heuristics based on French-word/English-word couples as used in (Okumura & Hovy, 1994; Rigau & Agirre, 1995; Atserias et al. 1997). These heuristics are being successfully used to build the Spanish and Basque WordNets included in the EuroWordNet project. Nevertheless, these word couples have also their drawbacks, since bilingual senses are not taken into account.

Thanks to the use of bilingual senses, WordNet and LPPL could be enriched with the supplementary information appearing in bilingual dictionaries, e.g. collocational information (Fontenelle, 1997).

At present, we are building the Basque WordNet, linked to the EuroWordNet and ITEM projects, making use of the techniques presented in chapter VI and the word couples that we have just mentioned applied to a Basque-English bilingual dictionary (Aulestia & White, 1982). The Spanish WordNet currently under construction, could be also fed into the Basque WordNet using a Basque-Spanish dictionary (Elhuyar, 1996). Using several bilingual dictionaries (Basque-Spanish, Basque-English and Spanish-English) coverage and precision could be improved.

The methods developed for this chapter can be used to join structured lexical resources in general, and this can have a heavy impact on the construction of future ontologies and LKBs. A given resource can be fed with the knowledge in another (in the same language or in a different one), and this looks like a promising avenue in the building of richer ontologies, following the proposals of the *ANSI Ad Hoc* committee on *Ontology Standards*[32] (Hovy, 1997a; 1997b).

X.C.4.b) Genus disambiguation

Although the obtained results are very good, there is still room for improvement. As proposed in a joint paper (Rigau et al. 1998), after applying the Genus disambiguation techniques (cf. chapter VI) on the DGILE (Alvar, 1987) Spanish dictionary, we clustered the genus according to the WordNet semantic code assigned. If only the senses appearing more frequently for each semantic code are considered, precision improves considerably, at the cost of coverage. We tried this method on LPPL too, but due to the small size of the dictionary, the frequencies were not high enough, and precision did not improve.

The research made in conjunction with the computational lexicography group in the Polytechnic University of Catalonia suggests that the developed method is successful with both small and large dictionaries. From larger dictionaries we get wider and more interesting hierarchies, offering also better choices for improvement.

Regarding the voting results, we think it would be interesting to analyze more sophisticated methods. In a small study, we observed that considering only decisions involving a majority of at least 5 heuristics, precision would rise up to 95%, but reducing coverage down to 18%.

On the other hand, when disambiguating, we just used the information in the definition itself. We also plan to disambiguate whole hierarchies. For instance, when disambiguating a given genus, we could bear in mind the hyponyms and hypernyms of each sense of the genus and the disambiguated hyponyms of the definiendum.

In the same way, after linking the disambiguated hierarchies to the top layer of WordNet, we can take advantage of the extra information and try to re-disambiguate the hierarchies.

X.C.4.c) Linking isolated hierarchies extracted from dictionaries

When building the hierarchies, we have not taken into account the synonymy relation. Most of the literature does not pay any attention to synonymy as extracted from dictionaries, but in the case of LPPL many definitions of nouns give just synonyms (the 20% of all word senses). Artola (1993), in the LKB extracted from LPPL, copied the extracted relations between synonyms, and it would be interesting to evaluate the impact of such a method in the disambiguated hierarchy. Other

[32] http://ksl-web.stanford.edu/onto-std/

approaches for the representation of synonymy, such as grouping all synonym word senses in a single concept (WordNet), would have to be studied too.

Although the method to link isolated hierarchies using the top layer of WordNet gave promising results, the quality of the obtained hierarchy was not thoroughly evaluated. At present, there is no agreed procedure to evaluate the quality of ontologies, apart from the number of correct hyponym/hypernim links, which we already provided (82%). This measure being very limited, it could be interesting to evaluate the method according to the usefulness for a given task, like information retrieval, for instance. Besides, we can not forget the *ANSI ad hoc Ontology Standards Group*, already mentioned, which is working also on ontology evaluation guidelines, without any published result for the time being.

X.C.4.d) Vicious circle

Among the three main tasks, that is, the LPPL-WordNet link, the disambiguation of genus in LPPL and the building of the top layer to connect the isolated hierarchies of LPPL, we have complex interrelations. In this dissertation, we have performed them sequentially, in the order just mentioned, but the interrelations among the three procedures would have to be better studied. Once the hierarchies of LPPL have been disambiguated and joined by means of the top layer of WordNet (via LPPL-WordNet links), we have more information to do the LPPL-WordNet link, as we are now linking full hierarchies, an21d better results can be expected. Besides, as mentioned above, after building the top layer, genus disambiguation would be easier. Moreover, with better bilingual links, both genus disambiguation and the top layer would improve. An iterative process suggests itself.

Another interesting approach could be the use of neural nets. All the results described in chapter VI –LPPL-WordNet link, the disambiguated hyponym/hypernim relations from LPPL, WordNet hierarchy– can be represented as an arch in a neural net. If we design an appropriate energy function, we can apply known techniques so as to find the optimal combination of arcs. Such a neural net would decide at the same time the best WordNet link and hypernym for a given LPPL sense.

X.C.4.e) Others

Even if we have studied the automatic construction and enrichment of LKBs, we have not explored all its implications. For instance, the **extraction from the *differentia*** in the definitions (Artola, 1993) was not touched. The use of the differentia has always been considered interesting and current work (see, for example, Richardson, 1997) shows a renewal of interest in this area.

Besides, we also think that the analysis of the example sentences can give complementary information, as they give interesting information about the context of the word sense.

The automatic building of multilingual hierarchies is a field close to this dissertation. When linking structured resources of different languages, we are implicitly building multilingual hierarchies. In fact, this involves studying whether it is possible to feed the information of ontologies in a given language (semi) automatically into another language. At the same time, questions arise such as whether we can build universal hierarchies, whether information from different hierarchies are compatible, whether it is convenient to link automatically the top layers, etc.

Regarding Basque, we have to mention the work carried out by our research group on the **Euskal Hiztegia** (Sarasola, 1997). The goal of this project is to extract a wide LKB for Basque, rich in semantic information. We have performed the study of the structure of the dictionary and translated following the TEI guidelines (Arriola et al. 1995; 1996a; 1996b). We have concluded the search of genus and special relators for noun definitions (Agirre et al. 1998), and are currently carrying out the analysis for verbs and adjectives, the analysis of example sentences, and the link to WordNet. Next, we plan to construct the disambiguated hierarchies for noun, verb and adjectives, following the method presented in chapter VI. Moreover, the study of the sublanguage used in the definitions of the Basque Dictionary is going on, and we will soon apply superficial syntactic techniques to extract further relations from the *differentia.*

Bibliography

Aduriz, I., Alegria, I., Artola, X., Ezeiza, N., Sarasola, K. and Urkia, M. 1997. A Spelling Corrector for Basque Based on morphology, in *Literary and Linguistic Computing*, vol. 12, no. 1. Oxford University Press (Oxford, England).

Agirre, E. 1993. Contribución de la Información Léxico-Sémantica en la Automatizción de la Corrección de Errores, in *Workshop sobre Lexicografía Computacional.* Unpublished paper (Donostia, Basque Country).

Agirre, E. and Rigau, G. 1995. A proposal for Word Sense Disambiguation using Conceptual Distance, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Agirre, E. and Rigau, G. 1996a. Word Sense Disambiguation using Conceptual Density, in *Proc. of COLING* (Copenhagen, Denmark).

Agirre, E. and Rigau, G. 1996b. An Experiment on Word Sense Disambiguation of the Brown Corpus using WordNet, in *MCCS-96-291*. Computing Research Laboratory (Las Cruces, New Mexico).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994a. A methodology for the extraction of semantic knowledge from dictionaries using phrasal patterns, in *Proc. of IBERAMIA. IV Congreso Iberoamericano de Inteligencia Artificial*, pp. 263-270. McGraw-Hill (Caracas, Venezuela).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994b. Conceptual Distance and Automatic Spelling Correction, in *Proc. of the Workshop on Computational Linguistics for Speech and Handwriting Recognition* (Leeds, England).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994c. Intelligent Dictionary Help Systems, in Brekke, M.; Andersen. I.; Dahl, T. and Myking, J. (eds.) *Applications and Implications of current LSP Research.* Fakbokforlaget (Norway).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994d. Lexical Knowledge Representation in an Intelligent Dictionary Help System, in *Proc. of COLING* (Kyoto, Japan).

Agirre, E., Arregi, X., Artola, X., Díaz De Ilarraza, A., Sarasola K. 1995. Lexical-Semantic Information and Automatic Correction of Spelling Errors, in K. Korta & J. M. Larrazabal (eds.) *Semantics And Pragmatics Of Natural Language: Logical And Computational Aspects*, no. 1. Ilcli Series (Donostia, Basque Contry).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. and Soroa, A. 1997. Constructing an Intelligent Dictionary Help System, in *Natural Language Engineering*. Cambridge University Press (Cambridge, England).

Agirre, E., Ansa, O., Arregi, X., Arriola, J.M., Díaz de Ilarraza, A., Lersundi, M., Soroa, A. and Urizar, R. 1998a. Extracción de relaciones semánticas mediante gramáticas de restricciones, in *Proc. of Sociedad Española para el Procesamiento del Lenguaje Natural* (Alicante, Spain).

Agirre, E., Gojenola, K., Sarasola, K. and Voutilainen, A. 1998b. Towards a Single Proposal in Spelling Correction, in *Proc. of the joint COLING and ACL meeting*.

Agirre, E., Gojenola, K., Sarasola, K. and Voutilainen, A. 1998c. Towards a Single Proposal in Spelling Correction, in *UPV/EHU-LSI TR 8-98*. UPV-EHU (Donostia, Basque Country).

Ahlswede, T.E. 1989. New technique for identifying relational structures in dictionary definitions, in U. Zernik (eds.) *Proc. of the 1st Intl. Lexical Acquisition Workshop*.

ALPAC 1966. Language and Machine: Computers in Translation and Linguistics. National Research Council (Washington, USA).

Alshawi, H. 1989. Analysing dictionary definitions, in B. Boguraev, T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*, pp. 153-169. Longman (New York, USA).

Alvar, M. (ed.) 1987. Diccionario General Ilustrado de la Lengua Española. Biblograf (Barcelona, Catalonia).

Amsler, R. A. 1981. Taxonomy for English Noun and Verbs, in *Proc. of the 19th Annual Meeting of the Association for Computational Linguistics*, pp. 133-138.

Arregi, X. 1995. Anhitz: itzulpenean laguntzeko hiztegi-sistema eleanitza, in *Ph.D. thesis*. UPV-EHU (Donostia, Basque Country).

Arriola, J.M and Soroa, A. 1996. Lexical Information Extraction for Basque, in *Student Conference in Computational Linguistics* (Montreal, Canada).

Arriola, J.M., Artola X., Soroa A 1995. Análisis automático del diccionario Hauta-Lanerako Euskal Hiztegia, in *Procesamiento del Lenguaje Natural*, no. 17, pp. 173-181. SEPLN (Bilbo, Basque Country).

Arriola, J.M., Artola X., Soroa A. 1996. Automatic extraction of lexical information from an ordinary dictionary, in *Proc. of EURALEX* (Göteborg, Sweden).

Artola, X. 1993. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza, in *Ph.D. thesis*. UPV-EHU (Donostia, Basque Country).

Atserias, J., Climent, S., Farreres, X., Rigau, G. and Rodríguez, H. 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Tchark, Bulgary).

Aulestia, G. and White, L. 1992. Euskara-ingelesa hiztegia. Elkar (Donostia, Basque Country).

Bar-Hillel, Y. 1960. Automatic Translation of Languages, in F. Alt, A. Donald Booth, and R.E. Meagher (eds.) *Advances in Computers*. Academic Press (New York, USA).

Basili, R., Della Rocca, M., Pazienza, M.T. and Velardi, P. 1995. Contexts and categories: tuning a general purpose classification to sublanguages, in *Proceedings of the Conference on Recent Advances on Natural Language Processing* (Tzigov Chark, Bulgary).

Basili, R. Della Rocca, M. and Pazienza, M.T. 1997. Towards a Bootstrapping Framework for Corpus Semantic Tagging, in *Proc. of the ACL-SIGLEX Workshop on Tagging text with Lexical Semantics: Why, What and How* (Washington, USA).

Bateman, J.A. 1990. Upper modeling: organizing knowledge for natural language processing, in *Proc. of 5th Intl. Workshop on Natural Language Generation* (Pittsburgh, USA).

Biblograf 1992. Diccionario Vox/Harrap's Esencial Español-Inglés. Biblograf (Barcelona, Catalonia).

Binot, J.L. and Jensen, K. 1987. A semantic expert using an online standard dictionary, in *Proc. of IJCAI.*

Bisson, G. 1995. Why and How to Define a Similarity Measure for Object-Based Representation Systems, in N.J.I. Mars (eds.) *Towards Very Large Knowledge Bases.* IOS Press.

Boguraev, B. and Briscoe, T. 1987. Large Lexicons for Natural Language Processing: Utilising the Grammar Coding System of LDOCE, in *Computational Linguistics*, vol. 13, no. 3-4.

Boguraev, B. and Briscoe, T. (eds.) 1989. Computational Lexicography for Natural Language Processing. Longman (New York, USA).

Briscoe, T., Copestake, A. and Boguraev, B. 1990. Enjoy the paper: lexical semantics via lexicology, in *Proc. of COLING.*

Briscoe, T., de Paiva, V. and Copestake, A. 1993. Inheritance, Defaults, and the Lexicon. Cambridge University Press (Cambridge, England).

Bruce, R. and Guthrie, L. 1991. Building a Noun Taxonomy from a Machine Readable Dictionary, in *MCCS-91-207.* Computing Research Laboratory (Las Cruces, New Mexico).

Bruce, R., Wilks, Y., Guthrie, L., Slator, B. and Dunning, T. 1992. NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour, in *MCCS-92-246.* Computing Research Laboratory (Las Cruces, New Mexico).

Byrd, R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S. and Rizk, O.A. 1987. Tools and Methods for Computational Lexicology, in *Computational Linguistics*, vol. 13, no. 2-4.

Byrd, R.J. 1990. Computational Lexicology for Building On-Line Dictionaries: the Wordsmith Experience, in L. Fignoni and C. Peters (eds.) *Computational Lexicology and Lexicography.* Giardini (Pisa, Italy).

Calzolari, N. 1983. Semantic links and the dictionary, in *Proc. of the Intl. Conference on Computers and the Humanities.*

Castellón, I. 1992. Lexicografia Computacional: Adquisición Automática de Conocimiento Léxico, in *Ph.D. thesis.* Universitat de Barcelona (Barcelona, Catalonia).

Chen, H., Lynch, K.J., Basu, K. and Ng, T.D. 1993. Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval, in *IEEE Expert.*

Chodorow, M.S., Byrd, R.J. and Heidorn, G.E. 1985. Extracting semantic hierarchies from large on-line dictionary, in *Proc. of the 23rd Annual Meeting of the Association for Computational Linguistics.*

Chodorow, M.S., Ravin, Y. and Sachar, H.E. 1988. A tool for investigating the synonymy relation in a sense desambiguated thesaurus, in *Proc. of the Conference on Applied Natural Language Processing* (Austin, USA).

Church, K. W., Hanks, P. 1990. Word Association Norms, Mutual Information, and Lexicography, in *Computational Linguistics*, vol. 16, no. 1.

Cohen, P. and Loiselle, C. 1988. Beyond ISA: Structures for Plausible Inference in Semantic Networks, in *Proc. of AAAI.*

Collins, A. M and Loftus, E. F. 1975. A Spreading-Activation Theory of Semantic processing, in *Psychological Review*, vol. 82, no. 6, pp. 407-428.

Copestake, A. 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary, in *Proc. of 1st Intl. Workshop on Inheritance in NLP* (Tilburg, Netherlands).

Cowie, J., Guthrie, J., and Guthrie, L. 1992. Lexical Disambiguation Using Simulated Annealing, in *Proc. of COLING* (Nantes, France), pp. 359-365.

Cucchiarelii, A. and Velardi, P. 1997. Automatic Selection of Class Labels from a Thesaurus for an Effective Tagging of Corpora, in *Proc. of the 5th Conference on Applied Natural Language Processing*, pp. 380-387.

Cunningham, H., Humphreys, K., Wilks, Y. and Gaizauskas, R. 1997. Software Infrastructure for Natural Language Processing, in *Proceedings of the Fifth Conference on Applied Natural Language Processing.*

Damerau, F.A. 1964. A technique for computer detection and correction of spelling errors, in *Information Processing and Management*, vol. 7, pp. 171-176.

Dietterich, T.G. 1997. Machine Learning Research: Four Current Directions, in *AI magazine*, vol. 18, no. 4, pp. 97-136.

EDR 1993. Electronic Dictionary Technical Guide, in *TR-042*. Electronic Dictionary Research Institute (Tokyo, Japan).

Elhuyar 1996. Elhuyar euskara-gaztelania hiztegia. Elhuyar K.E. (Usurbil, Basque Country).

Firth, J. 1956. A synopsis of linguistic theory 1930-1950, in M. Palmer (eds.) *Selected papers of J.R. Firth.* Longmans (London, England).

Fontenelle, T. 1997. Using a Bilingual Dictionary to Create Semantic Networks, in *International Journal of Lexicography*, vol. 10, no. 4.

Francis, S. and Kucera, H. 1967. Computing Analysis of Present-Day American english. Brown University Press.

Gale, W., Church, K. 1990. Poor Estimates of Context are Worse than none, in *Proc. of Compstat* (Dubrovnik, Yugoslavia). Springer-Verlag (New York, USA).

Gale, W., Church, K., Yarowsky, D. 1992. Work on Statistical Methods for Word Sense Disambiguation, in *Proc. of the AAAI Fall Symposium: Probabilistic Approaches to Natural Language Processing.*

Gale, W. A., Church, K.W. and Yarowsky, D. 1993. A Method for Disambiguating Word Senses in a Large Corpus, in *Computing and the Humanities*, no. 26, pp. 415-439.

Genthial, D., Courtin, J., Ménèzo, J. 1994. Towards a More User-Friendly Correction, in *Proc. of the Annual Meeting of the Association for Computational Linguistics* (Kyoto, Japan).

Golding, A. and Schaves, Y. 1996. Combining trigram-based and feature-based methods for context-sensitive spelling correction, in *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics* (Santa Cruz, USA).

Golding, A. R. 1995. A Bayesian hybrid method for context-sensitive spelling correction, in *Proc. of the 3rd Workshop on Very Large Corpora* (Cambridge, USA), pp. 39-53.

Gove, P.B. (ed.) 1969. The Webster's Seventh New Collegiate Dictionary. Merrian-Webster (Springfiled, Massachusets).

Grefenstette, G. 1992. Finding Semantic Similarity in Raw Text: the Deese Antonyms, in *Proc. of the AAAI Fall Symposium: Probabilistic Approaches to Natural Language Processing.*

Grefenstette, G. 1996. Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches, in Boguraev & Pustejovsky (eds.) *Corpus Processing for Lexical Acquisition*, ch. 11, pp. 213-225. MIT Press (Cambridge, Massachusetts).

Grishman, R. and Sterling, J. 1994. Generalizing Automatically Generated Selectional Patterns, in *Proc. of the Annual Meeting of the Association for Computational Linguistics.*

Gruber, T.R. 1993. Towards Principles for the Design of Ontologies for Knowledge Sharing, in *Proc. of the Intl. Workshop on Formal Ontology* (Padova, Italy). also as Technical Report KSL 93-04 (Stanford University, USA).

Guarino, N. 1997. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration, in Pazienza, M.T. (ed.) *Information Extraction.* Springer (Berlin, Germany).

Hearst, M., Schütze, H. 1993. Customizing a Lexicon to Better Suit a Computational Task, in *Proc. of the Workshop on Extracting Lexical Knowledge.*

Hearst, M. 1991. Toward Noun Homonym Disambiguation Using Local Context in Large Text Corpora, in *Proc. of the 7th Annual Conference of the UW Centre for the New OED and Text Research* (Waterloo, Canada).

Helmreich, S., Guthrie, L. and Wilks, Y. 1993. The use of machine readable dictionaries in the Pangloss project, in *Proc. of the AAAI Spring Symposium on Buildings Lexicons for Machine Translation*. AAAI Press.

Heylen, D., Maxwell, K.G. and Armstrong-Warwick, S. 1993. Collocations, Dictionaries and MT, in *Proc. of the AAAI Spring Symposium on Buildings Lexicons for Machine Translation*. AAAI Press.

Hirst G. 1987. Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press (Cambridge, England).

Hobbs, J. 1985. Ontological Promiscuity, in *Proc. of the 23rd Annual Meeting of the Association for Computational Linguistics*.

Hornby, A.S. (ed.) 1974. Oxford Advanced Learner's Dictionary of Current English. Oxford University Press (Oxford, England).

Hovy, E. and Nirenburg, S. 1992. Approximating an Interlingua in a Principled Way, in *Proceedings of the DARPA Speech and Natural Language Workshop* (Arden House, NY.).

Hovy, E. 1997a. Constructing and Using Large Ontologies, in *Unpublished Presentation on the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Ressources for NLP Applications* (Madrid, Spain).

Hovy, E. 1997b. A Standard for Large Ontologies, in *NSF Workshop on R&D Opportunities in the Government* (Waxhington, USA).

Ide, N. and Véronis, J. 1998. Introduction to the Special Issue on Word Sense Desambiguation: The State of the Art, in *Computational Linguistics*, vol. 24, no. 1.

Ide, N. and Véronis, J. 1994. Extracting Knowledge Bases From Machine-Readable Dictionaries: Have We Wasted Our Time?, in K. Fuchi and T. Yokoi (eds.) *Knowledge Building and Knowledge Sharing*. Ohmsha, Ltd. and IOS Press.

Ingels, P. 1996. Connected Text Recognition Using Layered HMMs and Token Passing, in K. Oflazer and H. Somers (eds.) *Proc. of the 2nd Conference on New Methods in Language Processing*, pp. 121-132.

Ingels, P. 1997. A Robust Text Processing Technique Applied to Lexical Error Recovery, in *Ph.D. thesis*. Department of Conputer and Information Science (Linköping, Sweden).

Ispell 1993. International Ispell Version 3.1.00.

Jones, M. P. and Martin, J. H. 1997. Contextual Spelling Correction Using Latent Semantic Analysis, in *Proc. of the Conference on Applied Natural Language Processing*, pp. 166-173.

Karlsson, F., Voutilainen, A., Heikkila, J. and Anttila, A. 1995. Constrait Grammar: a Language Independent System for Parsing Unrstricted Text. Mouton de Gruyter.

Karov, Y. and Edelman, S. 1996. Learning Similaity-Based Word Sense Disambiguation From Sparse Data, in *Proc. of the 6th Workshop on Very Large Corpora* (Copenhagen, Denmark).

Karov, Y. and Edelman, S. 1998. Similarity-based Word Sense Disambiguation, in *Computational Linguistics*, vol. 24, no. 1.

Kernighan, M., Church, K., Gale, W. 1990. A Spelling Program Based on a Noisy Channel Model, in *Proc. of COLING.*

Kilgarriff, A. 1997a. I don't believe in word senses, in *Computing and the Humanities*, no. 2.

Kilgarriff, A. 1997b. Evaluating Word Sense Disambiguation Programs: Progress Report, in *ITRI-97-11 Technical Report.* University of Brighton.

Kirkpatrick, B. 1987. Roget's Thesaurus. Longman (Harlow, England).

Klavans, J. and Tzoukermann, E. 1995. Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons, in *Machine Translation*, vol. 10, no. 3.

Knight, K. and Luk, S. 1994. Building a Large-Scale Knowledge Base for Machine Translation, in *Proc. of AAAI.*

Kozima, H. and Furugori, T. 1993. Similarity between Words Computed by Spreading Activation on an English Dictionary, in *Proc. of the 6th Conference of the European Chapter of the Association for Computational Linguistics.*

Kozima, H. and Ito, A. 1995. Context-Sensitive Measurement of Word Distance by Adaptive Scaling of a Semantic Space, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Kukich, K. 1990. A Comparison of Some Novel and Traditional Lexical Distance Metrics for Spelling Correction, in *Proc. of INNC* (Paris, France).

Kukich, K. 1992. Techniques for Automatically Correcting Words in Text, in *ACM Computing Surveys*, vol. 24, no. 4, pp. 377-439.

Larousse 1980. Le plus petit Larousse. Larousse (Paris, France).

Leacock, C., Chodorow, M. and Miller, G.A. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification, in *Computational Linguistics*, vol. 24, no. 2.

Lee, J.L. 1997. Similarity-Based Approaches to Natural Language Processing, in *Ph.D. thesis.* Harvard University Technical Report TR-11-97 (Cambridge, Massachusetts).

Lenat, D.B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure, in *Communications of the ACM*, vol. 38, no. 11.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, in *Proc. of the 1986 SIGDOC conference.* ACM (New York, USA).

Li, H. and Abe, N. 1995. Generalizing Case Frames Using a Thesaurus and the MDL Principle, in *Proc. of Recent Advances on Natural Language Processing.*

Li, H. and Abe, N. 1996. Learning Dependencies between Case Frame Slots, in *Proc. of the 13th Conference on Machine Learning.*

Mahesh, K., Nirenburg, S., Cowie, J. and Farwell, D. 1996. An Assesment of Cyc for Natural Language Processing, in *MCCS-96-302*. Computing Research Laboratory (Las Cruce, USA).

Mahesh, K., Nirenburg, S. and Beale, S. 1997. If You Have It, Flaunt It: Using Full Ontological Knowledge for Word Sense Disambiguation, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Maritxalar, M. and Díaz de Ilarraza, A. 1996. Hizkuntza baten ikaskuntza-prozesuan zeharreko tartehizkuntz osaketa, in *UPV/EHU-LSI TR 7-96*. EHUko Lengoaiak eta Sistema Informatikoak Saila (Donostia, Basque Country).

Markowitz, J. 1986. Semantically significant patterns in dictionary definitions, in *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Mays, E., Damerau, F., Mercer, R. 1991. Context Based Spelling Correction, in *Information Processing and Management*, vol. 27, no. 5.

McEnery, T. and Wilson, A. 1996. Corpus Linguistics. Edinburgh University Press.

McRoy, S. 1992. Using Multiple Knowledge Sources for Word Sense Discrimination, in *Computational Linguistics*, vol. 18, no. 1.

Menezo, J., Genthial D., and Courtin J. 1996. Reconnaisances pluri-lexicales dans CELINE, un système multi-agents de détection et correction des erreurs, in *Proc. of the Conference on NLP+IA* (Moncton, Canada).

Michiels, A. and Nöel, J. 1982. Approaches to thesaurus production, in *Proc. of COLING*.

Michiels, A. 1996. An experiment in translation selection and word sense discrimination, in *http://engdep1.philo.ulg.be/michiels/wdts.htm*.

Miller, G., Leacock, C., Tengi, R. and Bunker, T. 1993a. A Semantic Concordance, in *Proc. of ARPA Workshop on Human Language Technology*.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. Miller, K. and Tengi, R. 1993b. Five Papers on WordNet, in *CSL Report 43*. Cognitive Science Laboratory, Princeton University.

Morris, M. 1998. Ingelesa-euskara hiztegia. Eusenor (Donostia, Basque Country).

Nakamura, J., Nagao, M. 1988. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation, in *Proc. of COLING* (Budapest, Hungary).

Niwa, Y., Nitta, Y. 1994. Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, in *Proc. of COLING* (Kyoto, Japan).

Okumura, A. and Hovy, E. 1994. Lexicon-to-Ontology Concept Association Using a Bilingual Dictionary, in *Proc. of the 1st AMTA Conference*.

Onyshkevych, B. and Nirenburg, S. 1994. The Lexicon in the Scheme of KBMT Things, in *MCCS-94-277*. Computing Research Laboratory (Las Cruces, New Mexico).

OUP 1974. Oxford French-English Dictionary. Oxford University Press (Oxford, England).

Procter, P. (ed.) 1978. Longman Dictionary of Contemporary English. Longman (London).

Quillian, M. R. 1968. Semantic Memory, in *Ph.D. thesis*. Carnegie Institute of Technology.

Rada, R., Mili, H., Bicknell, E., Blettner, M. 1989. Development and Application of a Metric on Semantic Nets, in *IEEE Transactions on systems, man, and cybernetics*, vol. 19, no. 1.

Resnik, P. 1992. WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery, in *Proc. of AAAI*.

Resnik, P. 1993a. Semantic Classes and Syntactic Ambiguity, in *Proc. of the ARPA Workshop on Human Language Technology* (Princeton, USA).

Resnik, P. 1993b. Selection and Information: A Class-Based Approach to Lexical Relationships, in *Ph.D. thesis*. University of Pennsylvania.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in *Proc. of IJCAI*.

Resnik, P. 1997. Selectional Preference and Sense Disambiguation, in *Proc. of the ACL-SIGLEX Workshop on Tagging text with Lexical Semantics: Why, What and How* (Washington, USA).

Ribas, F. 1995. On Learning More Appropriate Selectional Restrictions, in *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*.

Richardson, S.D. 1997. Determining Similarity and Inferring Relations in a Lexical Knowledge Base, in *Ph.D. thesis*. The City University of New York.

Rigau, G. and Agirre, E. 1995. Disambiguating bilingual nominal entries against WordNet, in *Workshop On The Computational Lexicon - ESSLLI* (Barcelona, Catalonia).

Rigau, G., Rodríguez, H. and Turmo, J. 1995. Automatically Extracting Translation Links Using a Wide Coverage Semantic Taxonomy, in *Proc. of the 15th Intl. Conference on Artificial Intelligence* (Montpellier, France).

Rigau, G., Atserias, J. and Agirre, E. 1997. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation, in *Proc. of ACL/EACL* (Madrid, Spain).

Rigau, G., Rodriguez, H. and Agirre, E. 1998. Building Accurate Semantic Taxonomies from Monolingual MRDs, in *Proc. of the joint COLING and ACL meeting* (Montreal, Quebec).

Rigau, G. 1998. Automatic Acquisition of Lexical Knowledge from Machine Readable Dictionaries, in *Ph.D. thesis*. Polytechnic University of Catalonia (Barcelona, Catalonia).

Rizk, O. 1989. Sense Disambigution of Word Translations in Bilingual Dictionaries: Trying to Solve the Mapping Problem Automatically, in *RC 14666*. IBM Research Division, T.J. Watson Research Center (New York, USA).

Sarasola, I. 1997. Euskal Hiztegia. Gipuzkoako Kutxa (Donostia, Basque Country).

Schütze, H. 1998. Automatic Word Sense Discrimination, in *Computational Linguistics*, vol. 24, no. 1.

Schütze, H. 1992a. Word Sense Disambiguation With Sublexical Representations, in *Proc. of the AAAI Workshop on Statistically-Based Natural Language Processing Techniques.*

Schütze, H. 1992b. Context Space, in *Proc. of the AAAI Fall Symposium: Probabilistic Approaches to Natural Language Processing.*

Sinclair, J. (ed.) 1987. Colllins COBUILD English Language Dictionary. Collins (London, England).

Sussna, M. 1993. Word Sense Disambiguation for Free Text Indexing Using a Massive Semantic Network, in *Proc. of the 2nd Int. Conf. on Information and Knowledge Management* (Airlington, USA).

Svartvik, J. (ed.) 1990. The London-Lund Corpus of Spoken English. Lund University Press.

Towell, G. and Voorhees, E.M. 1998. Disambiguating Highly Ambiguous Words, in *Computational Linguistics*, vol. 24, no. 1.

Tsurumaru, H., Hitaka, T., and Yoshida, S. 1986. An attempt to automatic thesaurus construction from an ordinary japanese dictionary, in *Proc. of COLING.*

Tversky, A. 1977. Features of Similarity, in *Psychological Review*, vol. 84, no. 4, pp. 327-354.

Urkia, M. and Sagarna, A. 1990. Terminologia y lexicografia asistidas por ordenador: la experiencia de UZEI, in *Proc. of SEPLN* (Donostia, Basque Country).

Utiyama, M. and Hasida, K. 1997. Bottom-up alingnment of Ontologies, in *Proc. of the IJCAI Workshop on Ontologies and Multilingual Natural Language Processing.*

UZEI lantaldea 1982. Hizkuntzalaritza/1 hiztegia. UZEI (Donostia, Basque Country).

Véronis, J. and Ide N. 1990. Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries, in *Proc. of COLING* (Helsinki, Finland), vol. 2.

Vosse, T. 1992. Detecting and Correcting Morpho-syntactic Errors in Real Texts, in *Proc. of the 3rd Conference on Applied Natural Language Processing* (Trento, Italy), pp. 111-118.

Vosse, T. 1994. The Word Connection: Grammar-based Spelling Error Correction in Dutch, in *Ph.D. thesis.* Unit for Experimental and Theoretical Psychology (Univ. of Leiden, Holland).

Vossen, P. and Serail, I. 1990. Devil: a taxonomy-browser for decompositiona via the lexicon, in *Technical Report.* Faculty of Arts, University of Amsterdam.

Vossen, P., Díez-Orzas, P. and Peters, W. 1997. The Multilingual design of the EuroWordNet Database, in *Proc. of the IJCAI Workshop on Multilingual Ontologies for NLP Applications.*

Vossen, P. 1989. The structure of lexical knowledge as envisaged in the LINKS-project, in J-. Conolly and S. Dik (eds.) *Functional Grammar and the Computer.* Dordrecht: Foris.

Vossen, P. 1990. The end of the chain: Where does decomposition of lexical knowledge lead us eventually?, in *Proc. of the Conference on Functional Grammar.*

Vossen, P. 1996. Right or Wrong: combining Lexical Resources in the EuroWordNet Project, in *Proc. of EURALEX.*

Wilks, Y., Fass, D., Guo, C., McDonald, J.E., Plate, T., and Slator, B.M. 1990. Providing Machine Tractable Dictionary Tools, in *Machine Translation*, no. 5, pp. 99-154.

Wilks, Y., Slator, B.M., and Guthrie, L. 1996. Electric Words: Dictionaries, Computers, and Meanings. The MIT Press (Cambridge, USA).

Microsoft Corporation 1997. Word 97.

Yarowsky, D. 1992. Word sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, in *Proc. of COLING* (Nantes, France), pp. 454-460.

Yarowsky, D. 1993. One Sense per Collocation, in *Proc. of the 5th DARPA Speech and Natural Language Workshop.*

Yarowsky, D. 1994. Decision Lists for Lexical Ambiguity Resolution, in *Proc. of the Annual Meeting of the Association for Computational Linguistics.*

Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in *Proc. of the Annual Meeting of the Association for Computational Linguistics.*

Yokoi, T. 1995. The EDR Electronic Dictionary, in *Communications of the ACM*, vol. 38, no. 11.

Appendix A.

Code	Erref	Title	Chapters
A.1	Agirre & Rigau, 1995	A proposal for Word Sense Disambiguation using Conceptual Distance	III and IV
A.2	Agirre & Rigau, 1996a	Word Sense Disambiguation using Conceptual Density	III and IV
A.3	Agirre & Rigau, 1996b	An Experiment on Word Sense Disambiguation of the Brown corpus using WordNet	III and IV

Agirre, E. and Rigau, G. 1995. A proposal for Word Sense Disambiguation using Conceptual Distance, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Agirre, E. and Rigau, G. 1996a. Word Sense Disambiguation using Conceptual Density, in *Proc. of COLING* (Copenhagen, Denmark).

Agirre, E. and Rigau, G. 1996b. An Experiment on Word Sense Disambiguation of the Brown Corpus using WordNet, in *MCCS-96-291*. Computing Research Laboratory (Las Cruces, New Mexico).

N.B. In the postscript version of the thesis, all the papers can be obtained separately as a single compressed file from the home page of the author (*http://www.ji.si.ehu/users/eneko*) or directly from the following address *http://ixa.si.ehu.es/dokument/tesiak/EnekoThesisPapers.tar.gz*

Appendix B.

Code	Erref	Title	Chapters
B.1	Agirre et al., 1994b	Conceptual Distance and Automatic Spelling Correction	III and V
B.2	Agirre et al., 1995	Lexical-Semantic Information and Automatic Correction of Spelling Errors	V
B.3	Agirre et al., 1998b	Towards a Single Proposal in Spelling Correction	V
B.4	Agirre et al., 1998c	Towards a Single Proposal in Spelling Correction	V

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994b. Conceptual Distance and Automatic Spelling Correction, in *Proc. of the Workshop on Computational Linguistics for Speech and Handwriting Recognition* (Leeds, England).

Agirre, E., Arregi, X., Artola, X., Díaz De Ilarraza, A., Sarasola K. 1995. Lexical-Semantic Information and Automatic Correction of Spelling Errors, in K. Korta & J. M. Larrazabal (eds.) *Semantics And Pragmatics Of Natural Language: Logical And Computational Aspects*, no. 1. Ilcli Series (Donostia, Basque Contry).

Agirre, E., Gojenola, K., Sarasola, K. and Voutilainen, A. 1998b. Towards a Single Proposal in Spelling Correction, in *Proc. of the joint COLING and ACL meeting.*

Agirre, E., Gojenola, K., Sarasola, K. and Voutilainen, A. 1998c. Towards a Single Proposal in Spelling Correction, in *UPV/EHU-LSI TR 8-98*. UPV-EHU (Donostia, Basque Country).

N.B. In the postscript version of the thesis, all the papers can be obtained separately as a single compressed file from the home page of the author (*http://www.ji.si.ehu/users/eneko*) or directly from the following address *http://ixa.si.ehu.es/dokument/tesiak/EnekoThesisPapers.tar.gz*

Appendix C.

Code	Erref	Title	Chapters
C.1	Rigau & Agirre, 1995	Disambiguating bilingual nominal entries against WordNet	VI
C.2	Rigau et al., 1997	Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation	VI

Rigau, G. and Agirre, E. 1995. Disambiguating bilingual nominal entries against WordNet, in *Workshop On The Computational Lexicon - ESSLLI* (Barcelona, Catalonia).

Rigau, G., Atserias, J. and Agirre, E. 1997. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation, in *Proc. of ACL/EACL* (Madrid, Spain).

N.B. In the postscript version of the thesis, all the papers can be obtained separately as a single compressed file from the home page of the author (*http://www.ji.si.ehu/users/eneko*) or directly from the following address *http://ixa.si.ehu.es/dokument/tesiak/EnekoThesisPapers.tar.gz*

A Proposal for Word Sense Disambiguation using Conceptual Distance

Eneko Agirre.*
Lengoaia eta Sistema Informatikoak saila.
Euskal Herriko Unibertsitatea.
p.k. 649, 20080 Donostia. Spain. jibagbee@si.ehu.es

German Rigau.**
Departament de Llenguatges i Sistemes Informàtics.
Universitat Politècnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. g.rigau@lsi.upc.es

Abstract.

This paper presents a method for the resolution of lexical ambiguity and its automatic evaluation over the Brown Corpus. The method relies on the use of the wide-coverage noun taxonomy of WordNet and the notion of conceptual distance among concepts, captured by a Conceptual Density formula developed for this purpose. This fully automatic method requires no hand coding of lexical entries, hand tagging of text nor any kind of training process. The results of the experiment have been automatically evaluated against SemCor, the sense-tagged version of the Brown Corpus.

Keywords: Word Sense Disambiguation, Conceptual Distance, WordNet, SemCor.

1 Introduction

Word sense disambiguation is a long-standing problem in Computational Linguistics. Much of recent work in lexical ambiguity resolution offers the prospect that a disambiguation system might be able to receive as input unrestricted text and tag each word with the most likely sense with fairly reasonable accuracy and efficiency. The most extended approach is to attempt to use the context of the word to be disambiguated together with information about each of its word senses to solve this problem.

Several interesting experiments have been performed in recent years using preexisting lexical knowledge resources. (Cowie et al. 92) describe a method for lexical disambiguation of text using the definitions in the machine-readable version of the LDOCE dictionary as in the method described in (Lesk 86), but using simulated annealing for efficiency reasons. (Yarowsky 92) combines the use of the Grolier encyclopaedia as a training corpus with the categories of the Roget's International Thesaurus to create a statistical model for the word sense disambiguation problem with excellent results. (Wilks et al. 93) perform several interesting statistical disambiguation experiments using coocurrence data collected from LDOCE. (Sussna 93), (Voorhees 93), (Richarson et al. 94) define a disambiguation programs based in WordNet with the goal of improving precision and coverage during document indexing.

Although each of these techniques looks somewhat promising for disambiguation, either they have been only applied to a small number of words, a few sentences or not in a public domain corpus. For this reason we have tried to disambiguate all the nouns from real texts in the public domain sense tagged version of the Brown corpus (Francis & Kucera 67), (Miller et al. 93), also called Semantic Concordance or Semcor for short. We also use a public domain lexical knowledge source, WordNet (Miller 90). The advantage of this approach is clear, as Semcor provides an appropriate environment for testing our procedures in a fully automatic way.

* *Eneko Agirre was supported by a grant from the Basque Government.*
** *German Rigau was supported by a grant from the Ministerio de Educación y Ciencia.*

This paper presents a general automatic decision procedure for lexical ambiguity resolution based on a formula of the conceptual distance among concepts: Conceptual Density. The system needs to know how words are clustered in semantic classes, and how semantic classes are hierarchically organised. For this purpose, we have used a broad semantic taxonomy for English, WordNet. Given a piece of text from the Brown Corpus, our system tries to resolve the lexical ambiguity of nouns by finding the combination of senses from a set of contiguous nouns that maximises the total Conceptual Density among senses.

Even if this technique is presented as stand-alone, it is our belief, following the ideas of (McRoy 92) that full-fledged lexical ambiguity resolution should combine several information sources. Conceptual Density might be only one evidence of the plausibility of a certain word sense.

Following this introduction, section 2 presents the semantic knowledge sources used by the system. Section 3 is devoted to the definition of Conceptual Density. Section 4 shows the disambiguation algorithm used in the experiment. In section 5, we explain and evaluate the performed experiment. In section 6, we present further work and finally in the last section some conclusions are drawn.

2 WordNet and the Semantic Concordance

Sense is not a well defined concept and often has subtle distinctions in topic, register, dialect, collocation, part of speech, etc. For the purpose of this study, we take as the senses of a word those ones present in WordNet 1.4. WordNet is an on-line lexicon based on psycholinguistic theories (Miller 90). It comprises nouns, verbs, adjectives and adverbs, organised in terms of their meanings around semantic relations, which include among others, synonymy and antonymy, hypernymy and hyponymy, meronymy and holonymy. Lexicalised concepts, represented as sets of synonyms called synsets, are the basic elements of WordNet. The senses of a word are represented by synsets, one for each word sense. The version used in this work, WordNet 1.4, contains 83,800 words, 63,300 synsets (word senses) and 87,600 links between concepts.

The nominal part of WordNet can be viewed as a tangled hierarchy of hypo/ hypernymy relations. Nominal relations include also three kinds of meronymic relations, which can be paraphrased as member-of, made-of and component-part-of.

SemCor (Miller et al. 93) is a corpus where a single part of speech tag and a single word sense tag (which corresponds to a WordNet synset) have been included for all open-class words. SemCor is a subset taken from the Brown Corpus (Francis & Kucera, 67) which comprises approximately 250,000 words out of a total of 1 million words. The coverage in WordNet of the senses for open-class words in SemCor reaches 96% according to the authors. The tagging was done manually, and the error rate measured by the authors is around 10% for polysemous words.

3 Conceptual Density and Word Sense Disambiguation

A measure of the relatedness among concepts can be a valuable prediction knowledge source to several decisions in Natural Language Processing. For example, the relatedness of a certain word-sense to the context allows us to select that sense over the others, and actually disambiguate the word. Relatedness can be measured by a fine-grained conceptual distance (Miller & Teibel, 91) among concepts in a hierarchical semantic net such as WordNet. This measure would allow to discover reliably the lexical cohesion of a given set of words in English.

Conceptual distance tries to provide a basis for determining closeness in meaning among words, taking as reference a structured hierarchical net. Conceptual distance between two concepts is defined in (Rada et al. 89) as the length of the shortest path that connects the concepts in a hierarchical semantic net. In a similar approach, (Sussna 93) employs the notion of conceptual distance between network nodes in order to improve precision during document indexing. Following these ideas, (Agirre et al. 94) describes a new conceptual distance formula for the automatic spelling correction problem and (Rigau 94), using this conceptual distance formula, presents a methodology to enrich dictionary senses with semantic tags extracted from WordNet.

The measure of conceptual distance among concepts we are looking for should be sensitive to:

- the length of the shortest path that connects the concepts involved.

• the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.

• the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.

• the measure should be independent of the number of concepts we are measuring.

We have experimented with several formulas that follow the four criteria presented above. Currently, we are working with the Conceptual Density formula, which compares areas of subhierarchies.

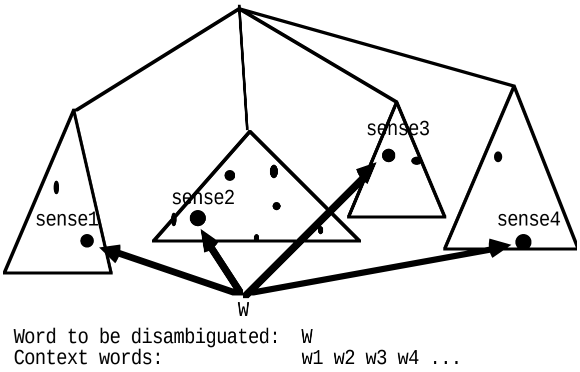


Figure 1: senses of a word in WordNet

As an example of how Conceptual Density can help to disambiguate a word, in figure 1 the word W has four senses and several context words. Each sense of the words belongs to a subhierachy of WordNet. The dots in the subhierarchies represent the senses of either the word to be disambiguated (W) or the words in the context. Conceptual Density will yield the highest density for the subhierarchy containing more senses of those, relative to the total amount of senses in the subhierarchy. The sense of W contained in the subhierarchy with highest Conceptual Density will be chosen as the sense disambiguating W in the given context. In figure 1, sense2 would be chosen.

Given a concept *c*, at the top of a subhierarchy, and given *nhyp* and *h* (mean number of hyponyms per node and height of the subhierarchy, respectively), the Conceptual Density for *c* when its subhierarchy contains a number *m* (marks) of senses of the words to disambiguate is given by the formula below:

$$\mathrm{CD(c,m)} = \frac{\sum_{i=0}^{m-1} \mathrm{nhyp}^{i}}{\sum_{i=0}^{h-1} \mathrm{nhyp}^{i}} \qquad \mathbf{(1)}$$

The numerator expresses the expected area for a subhierarchy containing *m* marks (senses of the words to be disambiguated), while the divisor is the actual area, that is, the formula gives the ratio between weighted marks below *c* and the number of descendant senses of concept *c*. In this way, formula 1 captures the relation between the weighted marks in the subhierarchy and the total area of the subhierarchy below *c*. The weight given to the marks tries to express that the height and the number of marks should be proportional.

nhyp is computed for each concept in WordNet in such a way as to satisfy equation 2, which expresses the relation among height, averaged number of hyponyms of each sense and total number of senses in a subhierarchy if it were homogeneous and regular:

$$\mathrm{descendants}_{c} = \sum_{i=0}^{h-1} \mathrm{nhyp}^{i} \qquad \mathbf{(2)}$$

Thus, if we had a concept *c* with a subhierarchy of height 5 and 31 descendants, equation 2 will hold that *nhyp* is 2 for *c*.

Conceptual Density weights the number of senses of the words to be disambiguated in order to make density equal to 1 when the number *m of* senses below *c* is equal to the height of the hierarchy *h*, to make density smaller than 1 if *m* is smaller than *h* and to make density bigger than 1 whenever *m* is bigger than *h*. The density can be kept constant for different *m*-s provided a certain proportion between the number of marks *m* and the height *h* of the subhierarchy is maintained. Both hierarchies **A** and **B** in figure 2, for instance, have Conceptual Density 1.

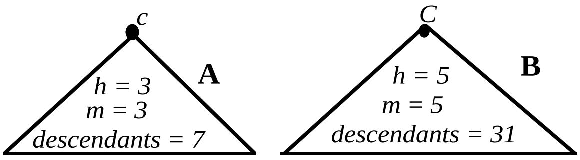


Figure 2: two hierarchies with CD = 1[1].

[1] *From formulas 1 and 2 we have:*

$$descendants(c) = 7 = \sum_{i=0}^{3-1} nhyp^{i} \Rightarrow nhyp = 2 \Rightarrow CD(c,3) = \sum_{i=0}^{3-1} 2^{i} \Big/ 7 = 7/7 = 1$$

$$descendants(c) = 31 = \sum_{i=0}^{5-1} nhyp^{i} \Rightarrow nhyp = 2 \Rightarrow CD(c,5) = \sum_{i=0}^{5-1} 2^{i} \Big/ 31 = 31/31 = 1$$

In order to tune the Conceptual Density formula, we have made several experiments adding two parameters, α and β. The a parameter modifies the strength of the exponential *i* in the numerator because *h* ranges between 1 and 16 (the maximum number of levels in WordNet) while *m* between 1 and the total number of senses in WordNet. Adding a constant b to *nhyp*, we tried to discover the role of the averaged number of hyponyms per concept. Formula 3 shows the resulting formula.

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} (nhyp+\beta)^{i^{\alpha}}}{descendants_c} \qquad \textbf{(3)}$$

After an extended number of runs which were automatically checked, the results showed that β does not affect the behaviour of the formula, a strong indication that this formula is not sensitive to constant variations in the number of hyponyms. On the contrary, different values of α affect the performance consistently, yielding the best results in those experiments with α near 0.20. The actual formula which was used in the experiments was thus the following:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c} \qquad \textbf{(4)}$$

4 The Disambiguation Algorithm Using Conceptual Density

Given a window size, the program moves the window one word at a time from the beginning of the document towards its end, disambiguating in each step the word in the middle of the window and considering the other words in the window as context.

The algorithm to disambiguate a given word w in the middle of a window of words W roughly proceeds as follows. First, the algorithm represents in a lattice the nouns present in the window, their senses and hypernyms (step 1). Then, the program computes the Conceptual Density of each concept in WordNet according to the senses it contains in its subhierarchy (step 2). It selects the concept c with highest density (step 3) and selects the senses below it as the correct senses for the respective words (step 4). If a word from W:

- has a single sense under c, it has already been disambiguated.
- has not such a sense, it is still ambiguous.
- has more than one such senses, we can eliminate all the other senses of w, but have not yet completely disambiguated w.

The algorithm proceeds then to compute the density for the remaining senses in the lattice, and continues to disambiguate words in W (back to steps 2, 3 and 4). When no further disambiguation is possible, the senses left for w are processed and the result is presented (step 5). To illustrate the process, consider the following text extracted from SemCor:

> *The jury(2) praised the administration(3) and operation(8) of the Atlanta Police_Department(1), the Fulton_Tax_Commissioner_'s_Office, the Bellwood and Alpharetta prison_farms(1), Grady_Hospital and the Fulton_Health_Department.*

Figure 3: sample sentence from SemCor

The underlined words are nouns represented in WordNet with the number of senses between brackets. The noun to be disambiguated in our example is *operation*., and a window size of five will be used.

(step 1) The following figure shows partially the lattice for the example sentence. As far as *Prison_farm* appears in a different hierarchy we do not show it in figure 4:

```
police_department_0
   => local department, department of
      local government
      => government department
         => department
         jury_1, panel
            => committee, commission
         operation_3, function
            => division
               => administrative unit
                  => unit
                     => organization
                        => social group
                           => people
                              => group

administration_1, governance...
jury_2
   => body
      => people
         => group, grouping
```

Figure 4: partial lattice for the sample sentence

The concepts in WordNet are represented as lists of synonyms. Word senses to be disambiguated are shown in bold. Underlined concepts are those selected with highest Conceptual Density. Monosemic nouns have sense number 0.

(Step 2) < administrative_unit>, for instance, has underneath 3 senses to be disambiguated and a subhierarchy size of 96 and therefore gets a Conceptual Density of 0.256. Meanwhile, <body>, with 2 senses and subhierarchy size of 86, gets 0.062.

(Step 3) <administrative_unit>, being the concept with highest Conceptual Density is selected.

(Step 4) **Operation_3**, **police_department_0** and **jury_1** are the senses chosen for *operation*, *Police Department* and *jury*. All the other concepts below <administrative_unit> are marked so that they are no longer selected. Other senses of those words are deleted from the lattice e.g. **jury_2**. In the next loop of the algorithm <body> will have only one disambiguation-word below it, and therefore its density will be much lower. At this point the algorithm detects that further disambiguation is not possible, and quits the loop.

(Step 5) The algorithm has disambiguated **operation_3**, **police_department_0**, **jury_1** and **prison_farm_0** (because this word is monosemous in WordNet), but the word *administration* is still ambiguous. The output of the algorithm , thus, will be that the sense for *operation* in this context, i.e. for this window, is **operation_3**. The disambiguation window will move rightwards, and the algorithm will try to disambiguate *Police Department* taking as context *administration*, *operation*, *prison farms* and whichever noun is first in the next sentence.

The disambiguation algorithm has and intermediate outcome between completely disambiguating a word or failing to do so. In some cases the algorithm returns several possible senses for a word. In this experiment we treat this cases as failure to disambiguate.

5 The Experiment

We selected one text from SemCor at random: br-a01 from the gender "Press: Reportage". This text is 2079 words long, and contains 564 nouns. Out of these, 100 were not found in WordNet. From the 464 nouns in WordNet, 149 are monosemous (32%).

The text plays both the role of input file (without semantic tags) and (tagged) test file. When it is treated as input file, we throw away all non-noun words, only leaving the lemmas of the nouns present in WordNet. The program does not face syntactic ambiguity, as the disambiguated part of speech information is in the input file. Multiple word entries are also available in the input file, as long as they are present in WordNet. Proper nouns have a similar treatment: we only consider those that can be found in WordNet. Figure 5 shows the way the algorithm would input the example sentence in figure 3 after stripping non-noun words.

After erasing the irrelevant information we get the words shown in figure 6[2].

The algorithm then produces a file with sense tags that can be compared automatically with the original file (c.f. figure 5).

```
<s>
<wd>jury</wd><sn>[noun.group.0]</sn><tag>NN</tag>
<wd>administration</wd><sn>[noun.act.0]</sn><tag>NN</tag>
<wd>operation</wd><sn>[noun.state.0]</sn><tag>NN</tag>
<wd>Police_Department</wd><sn>[noun.group.0]</sn><tag>NN</tag>
<wd>prison_farms</wd><mwd>prison_farm</mwd><msn>[noun.artifact.0]</msn><tag>NN</tag>
</s>
```

Figure 5: Semcor format

```
jury administration operation Police_Department prison_farm
```

Figure 6: input words

[2] *Note that we already have the knowledge that police department and prison farm are compound nouns, and that the lemma of prison farms is prison farm.*

Deciding the optimum context size for disambiguating using Conceptual Density is an important issue. One could assume that the more context there is, the better the disambiguation results would be. Our experiment shows that precision[3] increases for bigger windows, until it reaches window size 15, where it gets stabilised to start decreasing for sizes bigger than 25 (c.f. figure 7). Coverage over polysemous nouns behaves similarly, but with a more significant improvement. It tends to get its maximum over 80%, decreasing for window sizes bigger than 20.

Precision is given in terms of polysemous nouns only. The graphs are drawn against the size of the context[4] that was taken into account when disambiguating.

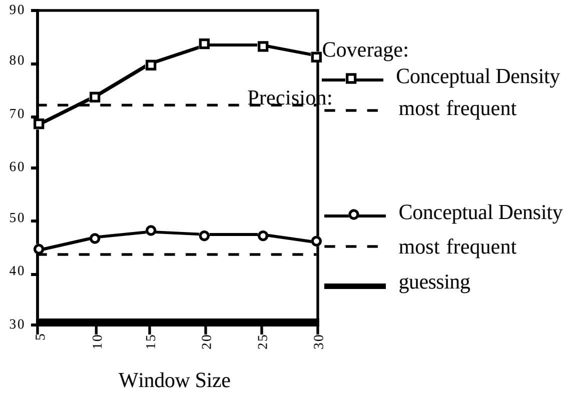


Figure 7: precision and coverage

Figure 7 also shows the guessing baseline, given when selecting senses at random. First, it was calculated analytically using the polysemy counts for the file, which gave 30% of precision. This result was checked experimentally running an algorithm ten times over the file, which confirmed the previous result.

We also compare the performance of our algorithm with that of the "most frequent" heuristic. The frequency counts for each sense were collected using the rest of SemCor, and then applied to the text. While the precision is similar to that of our algorithm, the coverage is nearly 10% worse.

All the data for the best window size can be seen in table 1. The precision and coverage shown in the preceding graph was for polysemous nouns only. If we also include monosemic nouns precision raises from 47.3% to 66.4%, and the coverage increases from 83.2% to 88.6%.

% w=25	Cover.	Prec.	Recall
polysemic	83.2	47.3	39.4
overall	88.6	66.4	58.8

Table 1: overall data for the best window size

6 Further Work

Senses in WordNet are organised in lexicographic files which can be roughly taken also as a semantic classification. If the senses of a given word that are from the same lexicographic file were collapsed, we would disambiguate at a level closer to the homograph level of disambiguation.

Another possibility we are currently considering is the inclusion of meronymic relations in the Semantic Density algorithm. The more semantic information the algorithm gathers the better performance it can be expected.

At the moment of writing this paper more extensive experiments which include other three texts from SemCor are under way. With these experiments we would like to evaluate the two improvements outlined above. Moreover, we would like to check the performance of other algorithms for conceptual distance on the same set of texts.

This methodology has been also used for disambiguating nominal entries of bilingual MRDs against WordNet (Rigau & Agirre 95).

7 Conclusion

The automatic method for the disambiguation of nouns presented in this paper is ready-usable in any general domain and on free-running text, given part of speech tags. It does not need any training and uses word sense tags from WordNet, an extensively used lexical data base.

[3] *Precision is defined as the ratio between correctly disambiguated senses and total number of answered senses. Coverage is given by the ratio between total number of answered senses and total number of senses. Recall is defined as the ratio between correctly disambiguated senses and total number of senses.*

[4] *Context size is given in terms of nouns.*

The algorithm is theoretically motivated and founded, and offers a general measure of the semantic relatedness for any number of nouns in a text.

In the experiment, the algorithm disambiguated one text (2079 words long) of SemCor, a subset of the Brown corpus. The results were obtained automatically comparing the tags in SemCor with those computed by the algorithm, which would allow the comparison with other disambiguation methods.

The results are promising, considering the difficulty of the task (free running text, large number of senses per word in WordNet), and the lack of any discourse structure of the texts.

Acknowledgements

We wish to thank all the staff at CRL in New Mexico State University, specially Jim Cowie, Joe Guthtrie, Louise Guthrie and David Farwell. We would also like to thank Ander Murua, who provided mathematical assistance, Xabier Arregi, Jose Mari Arriola, Xabier Artola, Arantxa Diaz de Ilarraza, Kepa Sarasola, and Aitor Soroa from the Computer Science Department of EHU and Francesc Ribas, Horacio Rodríguez and Alicia Ageno from the CS Department of UPC.

References

(Agirre et al. 94) Agirre E., Arregi X., Diaz de Ilarraza A. and Sarasola K., *Conceptual Distance and Automatic Spelling Correction.* in Workshop on Speech recognition and handwriting. Leeds, England. 1994.

(Cowie et al. 92) Cowie J., Guthrie J., Guthrie L., *Lexical Disambiguation using Simulated annealing,* in proceedings of DARPA WorkShop on Speech and Natural Language, 238-242, New York, February 1992.

(Francis & Kucera 67) Francis S. and Kucera H., *Computational analisys of present-day American English*, Providence, RI: Brown University Press, 1967.

(Lesk 86) Lesk M., *Automatic sense disambiguation: how to tell a pine cone from an ice cream cone*, in Proceeding of the 1986 SIGDOC Conference, Association for Computing Machinery, New York, 1986.

(McRoy 92) McRoy S., *Using Multiple Knowledge Sources for Word Sense Discrimination*. Computational Linguistics 18(1), March, 1992.

(Miller 90) Miller G., *Five papers on WordNet*, Special Issue of International Journal of Lexicogrphy 3(4). 1990.

(Miller & Teibel 91) Miller G. and Teibel D., *A proposal for Lexical Disambiguation*, in Proceedings of DARPA Speech and Natural Language Workshop, 395-399, Pacific Grave, California. February, 1991

(Miller et al. 93) Miller G. Leacock C., Randee T. and Bunker R. *A Semantic Concordance,* in proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey, March, 1993.

(Miller et al. 94) Miller G., Chodorow M., Landes S., Leacock C. and Thomas R., *Using a Semantic Concordance for sense Identification,* in proceedings of ARPA Workshop on Human Language Technology, 232-235, 1994.

(Rada et al. 89) Rada R., Mili H., Bicknell E. and Blettner M., *Development an Applicationof a Metric on Semantic Nets,* in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30. 1989.

(Richarson et al. 94) Richarson R., Smeaton A.F. and Murphy J., *Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words*, in Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland. 1994.

(Rigau 94) Rigau G., *An Experiment on Semantic Tagging of Dictionary Definitions,* in WorkShop “The Future of the Dictionary”. Uriage-les-Bains, France. October, 1994. Also as a research report LSI-95-31-R. Departament de Llenguatges i Sistemes Informàtics. UPC. Barcelona. June 1995.

(Rigau & Agirre 95) Rigau G., Agirre E., *Disambiguating bilingual nominal entries against WordNet*, Seventh European Summer School in Logic, Language and Information, ESSLLI'95, Barcelona, August 1995.

(Sussna 93) Sussna M., *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network,* in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia USA. 1993.

(Voorhees 93) Voorhees E. *Using WordNet to Disambiguate Word Senses for Text Retrival*, in Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA, June 1993.

(Wilks et al. 93) Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B., *Providing Machine Tractablle Dictionary Tools,* in Semantics and the Lexicon (Pustejowsky J. ed.), 341-401, 1993.

(Yarowsky 92) Yarowsky, D. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.* In Proceedings of the 15th International Conference on Computational Linguistics (Coling'92), Nantes, France. 1992.

Word Sense Disambiguation using Conceptual Density

Eneko Agirre*
Lengoaia eta Sistema Informatikoak saila. Euskal Herriko Universitatea.
p.k. 649, 200800 Donostia. Spain. jibagbee@si.heu.es

German Rigau**
Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. g.rigau@lsi.upc.es

Abstract.

This paper presents a method for the resolution of lexical ambiguity of nouns and its automatic evaluation over the Brown Corpus. The method relies on the use of the wide-coverage noun taxonomy of WordNet and the notion of conceptual distance among concepts, captured by a Conceptual Density formula developed for this purpose. This fully automatic method requires no hand coding of lexical entries, hand tagging of text nor any kind of training process. The results of the experiments have been automatically evaluated against SemCor, the sense-tagged version of the Brown Corpus.

1 Introduction

Much of recent work in lexical ambiguity resolution offers the prospect that a disambiguation system might be able to receive as input unrestricted text and tag each word with the most likely sense with fairly reasonable accuracy and efficiency. The most extended approach use the context of the word to be disambiguated together with information about each of its word senses to solve this problem.

Interesting experiments have been performed in recent years using preexisting lexical knowledge resources: [Cowie et al. 92], [Wilks et al. 93] with LDOCE, [Yarowsky 92] with Roget's International Thesaurus, and [Sussna 93], [Voorhees 93], [Richardson et al. 94], [Resnik 95] with WordNet.

Although each of these techniques looks promising for disambiguation, either they have been only applied to a small number of words, a few sentences or not in a public domain corpus. For this reason we have tried to disambiguate all the nouns from real texts in the public domain sense tagged version of the Brown corpus [Francis & Kucera 67], [Miller et al. 93], also called Semantic Concordance or SemCor for short[1]. The words in SemCor are tagged with word senses from WordNet, a broad semantic taxonomy for English [Miller 90][2]. Thus, SemCor provides an appropriate environment for testing our procedures and comparing among alternatives in a fully automatic way.

The automatic decision procedure for lexical ambiguity resolution presented in this paper is based on an elaboration of the conceptual distance among concepts: Conceptual Density [Agirre & Rigau 95]. The system needs to know how words are clustered in semantic classes, and how semantic classes are hierarchically organised. For this purpose, we have used WordNet. Our system tries to resolve the lexical ambiguity of nouns by finding the combination of senses from a set of contiguous nouns that maximises the Conceptual Density among senses.

The performance of the procedure was tested on four SemCor texts chosen at random. For comparison purposes two other approaches, [Sussna 93] and [Yarowsky 92], were also tried. The results show that our algorithm performs better on the test set.

Following this short introduction the Conceptual Density formula is presented. The main procedure to resolve lexical ambiguity of nouns using Conceptual Density is sketched on section 3. Section 4 describes extensively the experiments and its results. Finally, sections 5 and 6 deal with further work and conclusions.

*Eneko Agirre was supported by a grant from the Basque Goverment. Part of this work is included in projects 141226-TA248/95 of the Basque Country University and PI95-054 of the Basque Government.

**German Rigau was supported by a grant from the Ministerio de Educación y Ciencia.

[1] Semcor comprises approximately 250,000 words. The tagging was done manually, and the error rate measured by the authors is around 10% for polysemous words.

[2] The senses of a word are represented by synonym sets (or synsets), one for each word sense. The nominal part of WordNet can be viewed as a tangled hierarchy of hypo/hypernymy relations among synsets. Nominal relations include also three kinds of meronymic relations, which can be paraphrased as member-of, made-of and component-part-of. The version used in this work is WordNet 1.4, The coverage in WordNet of senses for open-class words in SemCor reaches 96% according to the authors.

2 Conceptual Density and Word Sense Disambiguation

Conceptual distance tries to provide a basis for measuring closeness in meaning among words, taking as reference a structured hierarchical net. Conceptual distance between two concepts is defined in [Rada et al. 89] as the length of the shortest path that connects the concepts in a hierarchical semantic net. In a similar approach, [Sussna 93] employs the notion of conceptual distance between network nodes in order to improve precision during document indexing. [Resnik 95] captures semantic similarity (closely related to conceptual distance) by means of the information content of the concepts in a hierarchical net. In general these approaches focus on nouns.

The measure of conceptual distance among concepts we are looking for should be sensitive to:

- the length of the shortest path that connects the concepts involved.
- the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.
- the measure should be independent of the number of concepts we are measuring.

We have experimented with several formulas that follow the four criteria presented above. The experiments reported here were performed using the Conceptual Density formula [Agirre & Rigau 95], which compares areas of subhierarchies.

To illustrate how Conceptual Density can help to disambiguate a word, in figure 1 the word W has four senses and several context words. Each sense of the words belongs to a subhierarchy of WordNet. The dots in the subhierarchies represent the senses of either the word to be disambiguated (W) or the words in the context. Conceptual Density will yield the highest density for the subhierarchy containing more senses of those, relative to the total amount of senses in the subhierarchy. The sense of W contained in the subhierarchy with highest Conceptual Density will be chosen as the sense disambiguating W in the given context. In figure 1, sense2 would be chosen.

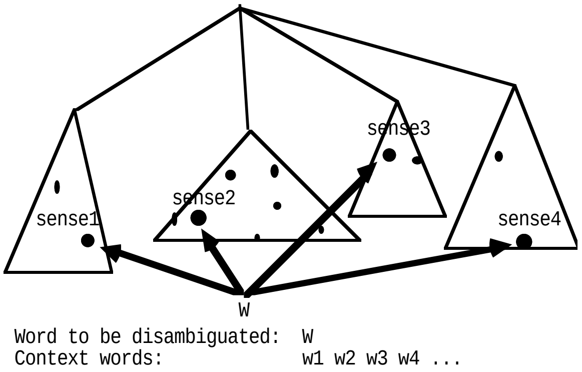


Figure 1: senses of a word in WordNet

Given a concept *c*, at the top of a subhierarchy, and given *nhyp* (mean number of hyponyms per node), the Conceptual Density for *c* when its subhierarchy contains a number *m* (marks) of senses of the words to disambiguate is given by the formula below:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c} \qquad \textbf{(1)}$$

Formula 1 shows a parameter that was computed experimentally. The 0.20 tries to smooth the exponential *i*, as *m* ranges between 1 and the total number of senses in WordNet. Several values were tried for the parameter, and it was found that the best performance was attained consistently when the parameter was near 0.20.

3 The Disambiguation Algorithm Using Conceptual Density

Given a window size, the program moves the window one noun at a time from the beginning of the document towards its end, disambiguating in each step the noun in the middle of the window and considering the other nouns in the window as context. Non-noun words are not taken into account.

The algorithm to disambiguate a given noun w in the middle of a window of nouns W (c.f. figure 2) roughly proceeds as follows:

```
(Step 1) tree := compute_tree(words_in_window)
         loop
(Step 2)   tree := compute_conceptual_distance(tree)
(Step 3)   concept := selecct_concept_with_highest_weigth(tree)
           if  concept = null then exitloop
(Step 4)   tree := mark_disambiguated_senses(tree,concept)
         endloop
(Step 5) output_disambiguation_result(tree)
```

Figure 2: algorithm for each window

First, the algorithm represents in a lattice the nouns present in the window, their senses and hypernyms (step 1). Then, the program computes the Conceptual Density of each concept in WordNet according to the senses it contains in its subhierarchy (step 2). It selects the concept c with highest Conceptual Density (step 3) and selects the senses below it as the correct senses for the respective words (step 4).

The algorithm proceeds then to compute the density for the remaining senses in the lattice, and continues to disambiguate the nouns left in W (back to steps 2, 3 and 4). When no further disambiguation is possible, the senses left for w are processed and the result is presented (step 5).

Besides completely disambiguating a word or failing to do so, in some cases the disambiguation algorithm returns several possible senses for a word. In the experiments we considered these partial outcomes as failure to disambiguate.

4 The Experiments

4.1 The texts

We selected four texts from SemCor at random: br-a01 (where a stands for gender "Press: Reportage"), br-b20 (b for "Press: Editorial"), br-j09 (j means "Learned: Science") and br-r05 (r for "Humour"). Table 1 shows some statistics for each text.

text	words	nouns	nouns in WN	monosemous
br-a01	2079	564	464	149 (32%)
br-ab20	2153	453	377	128 (34%)
br-j09	2495	620	586	205 (34%)
br-r05	2407	457	431	120 (27%)
total	9134	2094	1858	602 (32%)

Table 1: data for each text

An average of 11% of all nouns in these four texts were not found in WordNet. According to this data, the amount of monosemous nouns in these texts is bigger (32% average) than the one calculated for the open-class words from the whole SemCor (27.2% according to [Miller et al. 94]).

For our experiments, these texts play both the role of input files (without semantic tags) and (tagged) test files. When they are treated as input files, we throw away all non-noun words, only leaving the lemmas of the nouns present in WordNet.

4.2 Results and evaluation

One of the goals of the experiments was to decide among different variants of the Conceptual Density formula. Results are given averaging the results of the four files. Partial disambiguation is treated as failure to disambiguate. Precision (that is, the percentage of actual answers which were correct) and recall (that is, the percentage of possible answers which were correct) are given in terms of polysemous nouns only. Graphs are drawn against the size of the context[3] .

- **meronymy does not improve performance as expected.** A priori, the more relations are taken in account (e.i. meronymic relations, in addition to the hypo/hypernymy relation) the better density would capture semantic relatedness, and therefore better results can be expected.

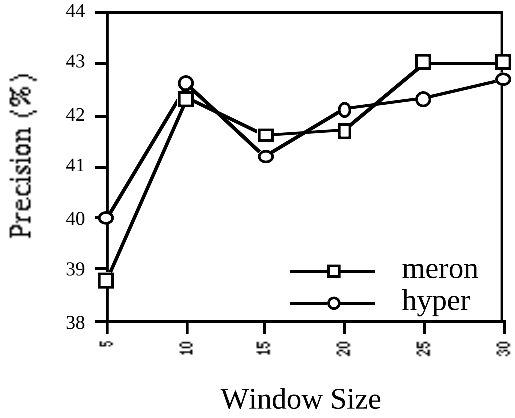


Figure 3: meronymy and hyperonymy

The experiments (see figure 3) showed that there is not much difference; adding meronymic information does not improve precision, and raises coverage only 3% (approximately). Nevertheless, in the rest of the results reported below, meronymy and hypernymy were used.

- **global nhyp is as good as local nhyp.** The average number of hyponyms or *nhyp* (c.f. formula 1) can be approximated in two ways. If an independent *nhyp* is computed for every concept in WordNet we call it *local nhyp*. If instead, a unique *nhyp* is computed using the whole hierarchy, we have *global nhyp*.

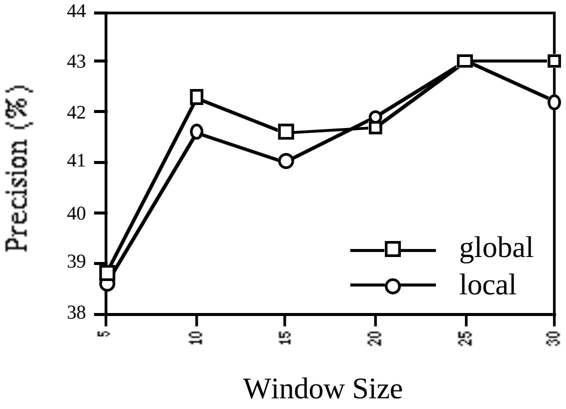


Figure 4: *local nhyp* vs. *global nhyp*

[3] context size is given in terms of nouns.

While *local nhyp* is the actual average for a given concept, *global nhyp* gives only an estimation. The results (c.f. figure 4) show that *local nhyp* performs only slightly better. Therefore *global nhyp* is favoured and was used in subsequent experiments.

• **context size: different behaviour for each text.** One could assume that the more context there is, the better the disambiguation results would be. Our experiments show that each file from SemCor has a different behaviour (c.f. figure 5) while br-b20 shows clear improvement for bigger window sizes, br-r05 gets a local maximum at a 10 size window, etc.

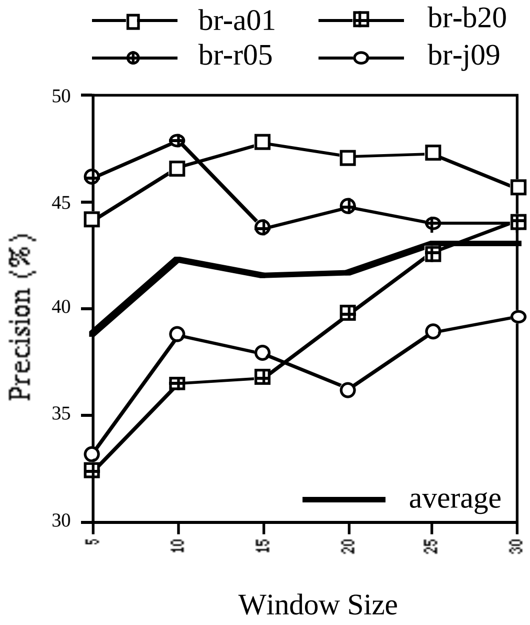


Figure 5: context size and different files

As each text is structured a list of sentences, lacking any indication of headings, sections, paragraph endings, text changes, etc. the program gathers the context without knowing whether the nouns actually occur in coherent pieces of text. This could account for the fact that in br-r05, composed mainly by short pieces of dialogues, the best results are for window size 10, the average size of this dialogue pieces. Likewise, the results for br-a01, which contains short journalistic texts, are best for window sizes from 15 to 25, decreasing significatly for size 30.

In addition, the actual nature of each text is for sure an important factor, difficult to measure, which could account for the different behavoiur on its own. In order to give an overall view of the performance, we consider the average behaviour.

• **file vs. sense.** WordNet groups noun senses in 24 lexicographer's files. The algorithm assigns a noun both an specific sense and a file label. Both file matches and sense matches are interesting to count. While the sense level gives a fine graded measure of the algorithm, the file level gives an indication of the performance if we were interested in a less sharp level of disambiguation. The granularity of the sense distinctions made in [Hearst, 91], [Yarowsky 92] and [Gale et al. 93] also called homographs in [Guthrie et al. 93], can be compared to that of the file level in WordNet.

For instance, in [Yarowsky 92] two homographs of the noun *bass* are considered, one characterised as MUSIC and the other as ANIMAL, INSECT. In WordNet, the 6 senses of *bass* related to music appear in the following files: ARTIFACT, ATTRIBUTE, COMMUNICATION and PERSON. The 3 senses related to animals appear in the files ANIMAL and FOOD. This means that while the homograph level in [Yarowsky 92] distinguishes two sets of senses, the file level in WordNet distinguishes six sets of senses, still finer in granularity.

Figure 6 shows that, as expected, file-level matches attain better performance (71.2% overall and 53.9% for polysemic nouns) than sense-level matches.

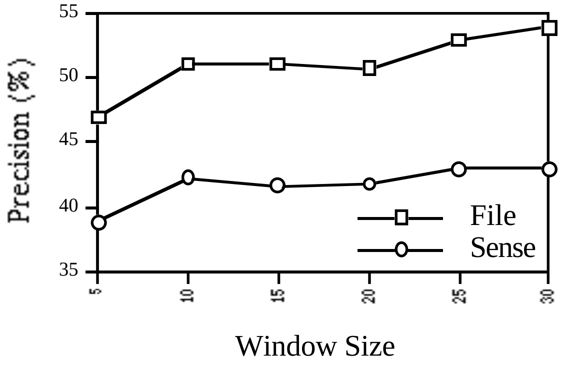


Figure 6: sense level vs. file level

• **evaluation of the results** Figure 7 shows that, overall, coverage over polysemous nouns increases significantly with the window size, without losing precision. Coverage tends to get stabilised near 80%, getting little improvement for window sizes bigger than 20.

The figure also shows the guessing baseline, given by selecting senses at random. This baseline was first calculated analytically and later checked experimentally. We also compare the performance of our algorithm with that of the "most frequent" heuristic. The frequency counts for each sense were collected using the rest of SemCor, and then applied to the four texts. While the precision is similar to that of our algorithm, the coverage is 8% worse.

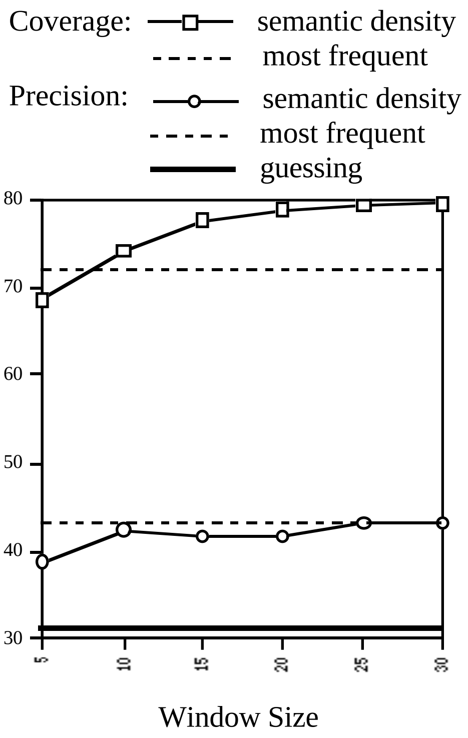


Figure 7: precision and coverage

All the data for the best window size can be seen in table 2. The precision and coverage shown in all the preceding graphs were relative to the polysemous nouns only. Including monosemic nouns precision raises, as shown in table 2, from 43% to 64.5%, and the coverage increases from 79.6% to 86.2%.

%	w=30	Cover.	Prec.	Recall
overall	File	86.2	71.2	61.4
	Sense		64.5	55.5
polysemic	File	79.6	53.9	42.8
	Sense		43	34.2

Table 2: overall data for the best window size

4.3 Comparison with other works

The raw results presented here seem to be poor when compared to those shown in [Hearst 91], [Gale et al. 93] and [Yarowsky 92]. We think that several factors make the comparison difficult. Most of those works focus in a selected set of a few words, generally with a couple of senses of very different meaning (coarse-grained distinctions), and for which their algorithm could gather enough evidence. On the contrary, we tested our method with **all** the nouns in a subset of an unrestricted public domain corpus (more than 9.000 words), making fine-grained distinctions among all the senses in WordNet.

An approach that uses hierarchical knowledge is that of [Resnik 95], which additionally uses the information content of each concept gathered from corpora. Unfortunately he applies his method on a different task, that of disambiguating sets of related nouns. The evaluation is done on a set of related nouns from Roget's Thesaurus tagged by hand. The fact that some senses were discarded because the human judged them not reliable makes comparison even more difficult.

In order to compare our approach we decided to implement [Yarowsky 92] and [Sussna 93], and test them on our texts. For [Yarowsky 92] we had to adapt it to work with WordNet. His method relies on cooccurrence data gathered on Roget's Thesaurus semantic categories. Instead, on our experiment we use saliency values[4] based on the lexicographic file tags in SemCor. The results for a window size of 50 nouns are those shown in table 3[5]. The precision attained by our algorithm is higher. To compare figures better consider the results in table 4, were the coverage of our algorithm was easily extended using the version presented below, increasing recall to 70.1%.

%	Cover.	Prec.	Recall
C.Density	86.2	71.2	61.4
Yarowsky	100.0	64.0	64.0

Table 3: comparison with [Yarowsky 92]

From the methods based on Conceptual Distance, [Sussna 93] is the most similar to ours. Sussna disambiguates several documents from a public corpus using WordNet. The test set was tagged by hand, allowing more than one correct senses for a single word. The method he uses has to overcome a combinatorial explosion[6] controlling the size of the window and “freezing” the senses for all the nouns preceding the noun to be disambiguated. In order to freeze the winning sense Sussna's algorithm is forced to make a unique choice. When Conceptual Distance is not able to choose a single sense, the algorithm chooses one at random.

Conceptual Density overcomes the combinatorial explosion extending the notion of conceptual distance from a pair of words to n words, and therefore can yield more than one correct sense for a word. For comparison, we altered our algorithm to also make random choices when unable to choose a single sense. We applied the algorithm Sussna considers best,

[4] We tried both mutual information and association ratio, and the later performed better.

[5] The results of our algorithm are those for window size 30, file matches and overall.

[6] In our replication of his experiment the mutual constraint for the first 10 nouns (the optimal window size according to his experiments) of file br-r05 had to deal with more than 200,000 synset pairs.

discarding the factors that do not affect performance significantly[7], and obtain the results in table 4.

%		Cover.	Prec.
C.Density	File	100.0	70.1
	Sense		60.1
Sussna	File	100.0	64.5
	Sense		52.3

Table 4: comparison with [Sussna 93]

A more thorough comparison with these methods could be desirable, but not possible in this paper for the sake of conciseness.

5 Further Work

We would like to have included in this paper a study on whether there is or not a correlation among correct and erroneous sense assignations and the degree of Conceptual Density, that is, the actual figure held by formula 1. If this was the case, the error rate could be further decreased setting a certain threshold for Conceptual Density values of winning senses. We would also like to evaluate the usefulness of partial disambiguation: decrease of ambiguity, number of times correct sense is among the chosen ones, etc.

There are some factors that could raise the performance of our algorithm:

•**Work on coherent chunks of text.** Unfortunately any information about discourse structure is absent in SemCor, apart from sentence endings The performance would gain from the fact that sentences from unrelated topics would not be considered in the disambiguation window.

• **Extend and improve the semantic data.** WordNet provides sinonymy, hypernymy and meronyny relations for nouns, but other relations are missing. For instance, WordNet lacks cross-categorial semantic relations, which could be very useful to extend the notion of Conceptual Density of nouns to Conceptual Density of words. Apart from extending the disambiguation to verbs, adjectives and adverbs, cross-categorial relations would allow to capture better the relations among senses and provide firmer grounds for disambiguating.

These other relations could be extracted from other knowledge sources, both corpus-based or MRD-based. If those relations could be given on WordNet senses, Conceptual Density could profit from them. It is our belief, following the ideas of [McRoy 92] that full-fledged lexical ambiguity resolution should combine several information sources. Conceptual Density might be only one of a number of complementary evidences of the plausibility of a certain word sense.

Furthermore, WordNet 1.4 is not a complete lexical database (current version is 1.5).

• **Tune the sense distinctions to the level best suited for the application.** On the one hand the sense distinctions made by WordNet 1.4 are not always satisfactory. On the other hand, our algorithm is not designed to work on the file level, e.g. if the sense level is unable to distinguish among two senses, the file level also fails, even if both senses were from the same file. If the senses were collapsed at the file level, the coverage and precision of the algorithm at the file level might be even better.

6 Conclusion

The automatic method for the disambiguation of nouns presented in this paper is ready-usable in any general domain and on free-running text, given part of speech tags. It does not need any training and uses word sense tags from WordNet, an extensively used lexical data base.

Conceptual Density has been used for other tasks apart from the disambiguation of free-running test. Its application for automatic spelling correction is outlined in [Agirre et al. 94]. It was also used on Computational Lexicography, enriching dictionary senses with semantic tags extracted from WordNet [Rigau 94], or linking bilingual dictionaries to WordNet [Rigau and Agirre 96].

In the experiments, the algorithm disambiguated four texts (about 10,000 words long) of SemCor, a subset of the Brown corpus. The results were obtained automatically comparing the tags in SemCor with those computed by the algorithm, which would allow the comparison with other disambiguation methods. Two other methods, [Sussna 93] and [Yarowsky 92], were also tried on the same texts, showing that our algorithm performs better.

Results are promising, considering the difficulty of the task (free running text, large number of senses per word in WordNet), and the lack of any discourse structure of the texts. Two types of results can be obtained: the specific sense or a coarser, file level, tag.

Acknowledgements

This work, partially described in [Agirre &Rigau 96], was started in the Computing Research Laboratory in New Mexico State University. We wish to thank all the staff of the CRL and specially Jim Cowie, Joe Guthtrie, Louise Guthrie and David Farwell. We would also like to thank Xabier Arregi, Jose mari Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Kepa Sarasola and Aitor Soroa from the Computer Science Faculty of EHU and Francesc Ribas, Horacio Rodríguez and Alicia Ageno from the Computer Science Department of UPC.

[7] Initial mutual constraint size is 10 and window size is 41. Meronymic links are also considered. All the links have the same weigth.

References

Agirre E., Arregi X., Diaz de Ilarraza A. and Sarasola K. 1994. *Conceptual Distance and Automatic Spelling Correction.* in Workshop on Speech recognition and handwriting. Leeds, England.

Agirre E., Rigau G. 1995. *A Proposal for Word Sense Disambiguation using conceptual Distance*, International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.

Agirre, E. and Rigau G. 1996. *An Experiment in Word SenseDisambiguation of the Brown Corpus Using WordNet*. Memoranda in Computer and Cognitive Science, MCCS-96-291, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Cowie J., Guthrie J., Guthrie L. 1992. *Lexical Disambiguation using Simulated annealing,* in proceedings of DARPA WorkShop on Speech and Natural Language, New York. 238-242.

Francis S. and Kucera H. 1967. *Computing analysis of present-day American English*, Providenc, RI: Brown University Press, 1967.

Gale W., Church K. and Yarowsky D. 1993. *A Method for Disambiguating Word Sense sin a Large Corpus*, in Computers and the Humanities, n. 26.

Guthrie L., Guthrie J. and Cowie J. 1993. *Resolving Lexical Ambiguity*, in Memoranda in Computer and Cognitive Science MCCS-93-260, Computing Research Laboratory, New Mexico State University. Las Cruces, New Mexico.

Hearst M. 1991. *Towards Noun Homonym Disambiguation Using Local Context in Large Text Corpora*, in Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research. Waterloo, Ontario.

McRoy S. 1992. *Using Multiple Knowledge Sources for Word Sense Discrimination*, Computational Linguistics, vol. 18, num. 1.

Miller G. 1990. *Five papers on WordNet*, Special Issue of International Journal of Lexicogrphy 3(4). 1990.

Miller G. Leacock C., Randee T. and Bunker R. 1993. *A Semantic Concordance*, in proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey.

Miller G., Chodorow M., Landes S., Leacock C. and Thomas R. 1994. *Using a Semantic Concordance for sense Identification,* in proceedings of ARPA Workshop on Human Language Technology, 232-235.

Rada R., Mili H., Bicknell E. and Blettner M. 1989. *Development an Applicationof a Metric on Semantic Nets*, in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30.

Resnik P. 1995. *Disambiguating Noun Groupings with Respect to WordNet Senses,* in Proceedings of the Third Workshop on Very Large Corpora, MIT.

Richardson R., Smeaton A.F. and Murphy J. 1994. *Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words*, in Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland.

Rigau G. 1994. *An experiment on Automatic Semantic Tagging of Dictionary Senses,* WorkShop "The Future of Dictionary", Aix-les-Bains, France. published as Research Report LSI-95-31-R. Computer Science Department. UPC. Barcelona.

Rigau G. and Agirre E. 1996. *Linking Bilingual Dictionaries to WordNet*, in proceedings of the 7th Euralex International Congress on Lexcography (Euralex'96), Gothenburg, Sweden, 1996.

Sussna M. 1993. *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*, in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia.

Voorhees E. 1993. *Using WordNet to Disambiguate Word Senses for Text Retrival*, in proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA.

Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. 1993. *Providing Machine Tractablle Dictionary Tools,* in Semantics and the Lexicon (Pustejovsky J. ed.), 341-401.

Yarowsky, D. 1992. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora,* in proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.

randa in Computer and Cognitive Science MCCS-93-260, Computing Research Laboratory, New Mexico State University. Las Cruces, New Mexico.

[Hearst, 91] Hearst M., 1991.*Towards Noun Homonym Disambiguation Using Local Context in Large Text Corpora*. Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research. Waterloo, Ontario.

[Lesk 86] Lesk M., 1986. *Automatic sense disambiguation: how to tell a pine cone from an ice cream cone*. Proceeding of the 1986 SIGDOC Conference, Association for Computing Machinery, New York.

[McRoy 92] McRoy S., 1992.*Using Multiple Knowledge Sources for Word Sense Discrimination*, Computational Linguistics, vol. 18, num. 1.

[Miller 90] Miller G., 1990. *Five papers on WordNet,* Special Issue of International Journal of Lexicography 3(4).

[Miller & Teibel 91] Miller G. and Teibel D., 1991. *A proposal for Lexical Disambiguation.* Proceedings of DARPA Speech and Natural Language Workshop, 395-399, Pacific Grove, California.

[Miller et al. 93] Miller G. Leacock C., Randee T. and Bunker R., 1993. *A Semantic Concordance*. Proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey.

[Miller et al. 94] Miller G., Chodorow M., Landes S., Leacock C. and Thomas R., 1994. *Using a Semantic Concordance for Sense Identification*. Proceedings of ARPA Workshop on Human Language Technology, 232-235.

[Rada et al. 89] Rada R., Mili H., Bicknell E. and Blettner M., 1989. *Development an Applicationof a Metric on Semantic Nets.* IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30.

[Resnik 93] Resnik P., 1993. *Semantic Classes and Syntactic Ambiguity.* Proceedings of ARPA Workshop on Human Language Technology, 303-308. Plainsboro, New Jersey.

[Resnik 95] Resnik P., 1995. *Disambiguating Noun Groupings with Respect to WordNet Senses.* Proceedings of the Third Workshop on Very Large Corpora, MIT. Cambridge, Massachusetts.

[Ribas 95] Ribas F., 1995. *On learning more Appropriate Selectional Restrictions*. Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, 112-118, Belfield, Dublin, Ireland.

[Richarson et al. 94] Richarson R., Smeaton A.F. and Murphy J., 1994. *Using WordNet as a Kownledge Base for Measuring Semantic Similarity between Words*. Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland.

[Rigau 94] Rigau G., 1994. *An experiment on Automantic Semantic Tagging of Dictionary Senses.* WorkShop "The Future of Dictionary", Aix-les-Bains, France.

[Rigau & Agirre 95] Rigau G., Agirre E., 1995. *Disambiguating bilingual nominal entries against WordNet*. Seventh European Summer School in Logic, Language and Information, ESSLLI'95, Barcelona.

[Schütze 92] Schütze H., 1992. *Context Space*. Workshop Notes of Fall Session of Statistically-Based Natural Language Processing Techniques, AAAI'92.

[Sussna 93] Sussna M., 1993.*Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network.* Proceedings of the Second International Conference on Information and Knowledge Management. Arlington, Virginia USA.

[Voorhees 93] Voorhees E., 1993.*Using WordNet to Disambiguate Word Senses for Text Retrival*. Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA.

[Wilks et al. 93] Wilks Y., Fass D., Guo C., McDonald J., Plate T. and Slator B., 1993. *Providing Machine Tractable Dictionary Tools.* Semantics and the Lexicon (Pustejovsky J. ed.), 341-401.

[Yarowsky 92] Yarowsky, D., 1992.*Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.* Proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.

[Yarowsky 93] Yarowsky, D., 1993. *One sense per Collocation*. Proceedings of ARPA Workshop on Human Language Technology, 266-271, Plainsboro, New Jersey.

- Compute the upper bound of this method using WordNet.

How correct this methodology can be? That is, words belonging to the same narrow context in SemCor can represent distant correct concepts in WordNet (having other incorrect ones closer).

7 Conclusion

The automatic method for the disambiguation of nouns presented in this paper is ready to use in any general domain, free-running text, given part of speech tags. It does not need any training and uses word sense tags from WordNet, a widely used lexical data base. The algorithm is theoretically motivated, and offers a general measure of the semantic relatedness for any number of nouns.

Conceptual Density has been used for other tasks apart from the disambiguation of free-running test. Its application for automatic spelling correction is outlined in [Agirre et al. 94]. It was also used on Computational Lexicography, enriching dictionary senses with semantic tags extracted from WordNet [Rigau 94], or linking bilingual dictionaries to WordNet [Rigau and Agirre 95].

In the experiments, the algorithm disambiguated four texts (more than 9,000 words long) of SemCor, a subset of the Brown corpus. The results were obtained automatically by comparing the tags in SemCor with those computed by the algorithm. This allows the comparison with other disambiguation methods. Two other methods, [Sussna 93] and [Yarowsky 92], were also tried on the same texts, showing that our algorithm performs better.

The results are promising, considering the difficulty of the task (free running text, large number of senses per word in WordNet), and the lack of any discourse structure of the texts. Two kinds of results can be obtained: the specific sense or a coarser, file level, tag.

Acknowledgements

We wish to thank all the staff of the CRL and specially Jim Cowie, Joe Guthtrie, Louise Guthrie and David Farwell. We would also like to thank Ander Murua, who provided mathematical assistance, Xabier Arregi, Jose Mari Arriola, Xabier Artola, Arantxa Diaz de Ilarraza, Kepa Sarasola, and Aitor Soroa from the Computer Science Department of EHU and Francesc Ribas, Horacio Rodríguez and Alicia Ageno from the Computer Science Department of UPC.

BIBLIOGRAPHY

[Agirre et al. 94] Agirre E., Arregi X., Diaz de Ilarraza A. and Sarasola K., 1994.*Conceptual Distance and Automatic Spelling Correction.* Workshop on Speech recognition and handwriting. Leeds, England.

[Church et al. 91] Church K., Gale W., Hanks P. and Hindle D., 1991.*Using Statistics in Lexical Analysis*. Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Zernik U. Ed. Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey.

[Cowie et al. 92] Cowie J., Guthrie J., Guthrie L., 1992. *Lexical Disambiguation using Simulated annealing.* Proceedings of DARPA WorkShop on Speech and Natural Language, 238-242, New York.

[Francis & Kucera 67] Francis S. and Kucera H., 1967.*Computational analisys of present-day American English*, Providence, RI: Brown University Press.

[Gale et al. 93] Gale W., Church K. and Yarowsky D., 1993. *A Method for Disambiguating Word Sense sin a Large Corpus*. Computers and the Humanities, n. 26.

[Guthrie et al. 93] Guthrie L., Guthrie J. and Cowie J., 1993. *Resolving Lexical Ambiguity*,. Memo-

%		Cover.	Prec.
C.Density	File	100.0	70.1
	Sense		60.1
Sussna	File	100.0	64.5
	Sense		52.3

Table 4: comparison with [Sussna 93]

6 Future Work

Initially, we would like to carry out a study on whether there is or is not a correlation between correct and erroneous sense assignations and the degree of Conceptual Density computed by formula 3. If this was the case, the error rate could be further decreased by setting a certain threshold for Conceptual Density values for winning senses.

There are other factors that could increase the performance of our algorithm:

- Work on coherent chunks of text.

Unfortunately any information about discourse structure is absent in SemCor, apart from sentence endings. If coherent pieces of discourse were taken as input, both performance and efficiency of the algorithm might improve. The performance would gain from the fact that sentences from unrelated topics would not be considered in the disambiguation window. We think that efficiency could also be improved if the algorithm worked on entire coherent chunks instead of one word at a time.

- Extend and improve the semantic data.

WordNet lacks cross-categorial semantic relations, which could be very useful for extending the notion of Conceptual Density of nouns to Conceptual Density of words. Apart from extending disambiguation to verbs, adjectives and adverbs, cross-categorial relations would allow the algorithm better capture the relations among senses and provide firmer grounds for disambiguating.

If Conceptual Density takes into account global relations among words, it may be advantageous to combine it with other sources of knowledge (both corpus-based or MRD-based) such as syntactic cues, word frequencies, collocations, selectional restrictions [Yarowsky 93], [Ribas 95], and so on. (c.f. [McRoy 92]). For instance, [Richardson et al. 94] defines conceptual similarity between two senses based on WordNet and informational measures taken from corpora, but does not give any evaluation of their method.

- Tune the sense distinctions to the level best suited for the application.

On the one hand, the sense distinctions made by WordNet 1.4 are not always satisfactory and, obviously, WordNet 1.4 is not a complete lexical Database. For instance, the three senses of abobe and the lack of connections among them, which are fixed up in WordNet 1.5. On the other hand, our algorithm is not designed to work on the file level, e.g. if the sense level is unable to distinguish among two senses, the file level also fails, even if both senses were from the same file. If the senses were collapsed at the file level, the coverage and precision of the algorithm at the file level might be better.

[12] *Initial mutual constraint size is 10 and window size is 41. Meronymic links are also considered. All the links have the same weigth.*

The raw results presented here seem to be poor when compared to those shown in [Hearst 91], [Gale et al. 93] and [Yarowsky 92]. We think that several factors make the comparison difficult. Most of those works focus on a selected set of a few words, generally with a couple of senses of very different meaning (coarse-grained distinctions), and for which their algorithm could gather enough evidence. On the contrary, we tested our method with **all** the nouns in a subset of an unrestricted public domain corpus (more than 9.000 words), making fine-grained distinctions among all the senses in WordNet.

[Guthrie et al. 93] tested their method in similar conditions to ours, but without performing an extensive and automatic testing. The results reported there seem to be lower than those shown here. In an experiment with 50 sample sentences from LDOCE, 47% of the words were correctly disambiguated to the sense level, and 72% to the homograph level (our file level would stand between their homograph and sense levels).

An approach that uses hierarchical knowledge is that of [Resnik 95], which additionally uses the information content of each concept gathered from corpora. Unfortunately he applies his method on a different task, that of disambiguating sets of related nouns. The evaluation is done on a set of related nouns from Roget's Thesaurus tagged by hand. The fact that some senses were discarded because the human judged them not reliable makes comparison even more difficult.

In order to compare our approach we decided to implement [Yarowsky 92] and [Sussna 93], and test them on our texts. For [Yarowsky 92] we had to adapt it to work with WordNet. His method relies on cooccurrence data gathered on Roget's Thesaurus semantic categories. Instead, on our experiment we use saliency values[9] based on the lexicographic file tags in SemCor (see Figure 4). The results for a window size of 50 are those shown in table 3[10]. The precision attained by our algorithm is higher. To compare figures better consider the results in table 4, were the coverage of our algorithm was easily extended using the version presented below, increasing recall to 70.1%.

%	Cover.	Prec.	Recall
C.Density	86.2	71.2	61.4
Yarowsky	100.0	64.0	64.0

Table 3: comparison with [Yarowsky 92]

From the methods based on Conceptual Distance, [Sussna 93] is the most similar to ours. Sussna disambiguates several documents from a public corpus using WordNet. The test set was tagged by hand, allowing more than one correct senses for a single word. The method he uses has to overcome a combinatorial explosion[11] controlling the size of the window and “freezing” the senses for all the nouns preceding the noun to be disambiguated. In order to freeze the winning sense Sussna's algorithm is forced to make a unique choice. When Conceptual Distance is not able to choose a single sense, he has to choose one at random.

Conceptual Density overcomes the combinatorial explosion extending the notion of conceptual distance from a pair of words to n words, and therefore can yield more than one correct sense for a word. For comparison, we altered our algorithm to also make random choices when unable to choose a single sense. We applied the algorithm Sussna considers best, discarding the factors that do not affect performance significantly[12], and obtain the results in table 4.

[9] *We tried both mutual information and association ratio, and the later performed better.*

[10] *The results of our algorithm are those for window size 30, file matches and overall.*

[11] *In our replication of his experiment the mutual constraint for the first 10 nouns (the optimal window size according to his experiments) of file br-r05 had to deal with more than 200.000 synset pairs.*

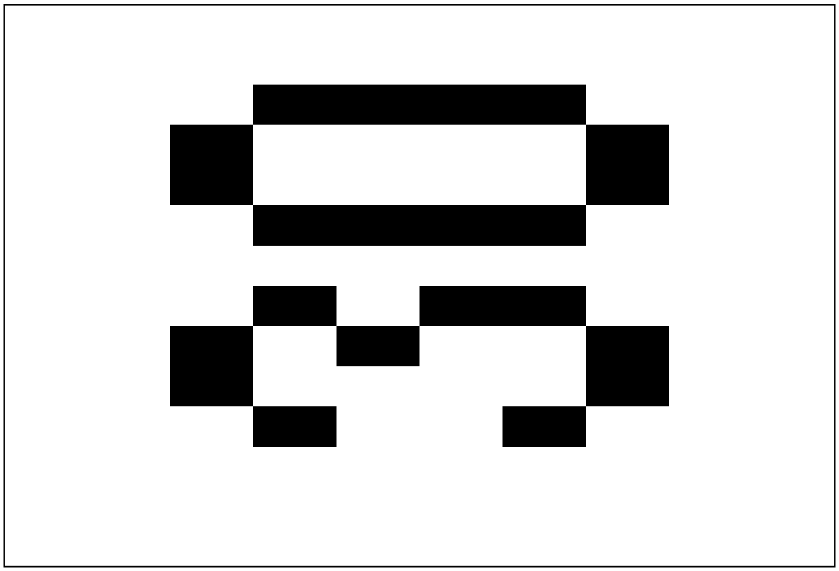


Figure 11: complete disambiguation and partial disambiguation

5.2.6 file vs. sense

WordNet synsets can be grouped by the lexicographic files they are coming from (e.g. ACT, ANIMAL, FOOD, etc.) Both file matches and synset matches are interesting to count. While the sense level gives a fine grained measure of the algorithm, the file level gives an indication of the performance if we were interested in a less precise level of disambiguation. The granularity of the sense distinctions made in [Hearst, 91], [Gale et al. 93] and [Yarowsky 92], also called homographs in [Guthrie et al. 93], can be compared to that of the file level in WordNet.

For instance, in [Yarowsky 92] two homographs of the noun bass are considered, one characterised as MUSIC and the other as ANIMAL, INSECT. In WordNet, the 6 senses of bass related to music appear in the following files: ARTIFACT, ATTRIBUTE, COMMUNICATION and PERSON. The 3 senses related to animals appear in the files ANIMAL and FOOD. This means that while the homograph level in [Yarowsky 92] distinguishes two sets of senses, the file level in WordNet distinguishes six sets of senses, still finer in granularity.

The following figure shows that, as expected, file-level matches attain better performance (71.2% overall and 53.9% for polysemic nouns) than sense-level matches.

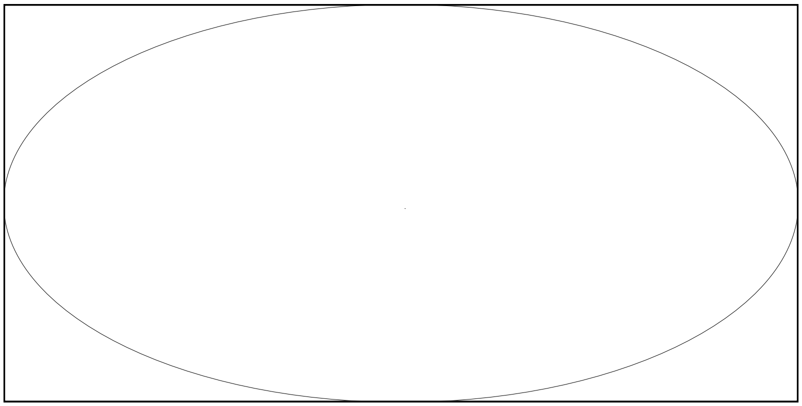


Figure 12: sense level v. file level

5.3 Comparison with other works

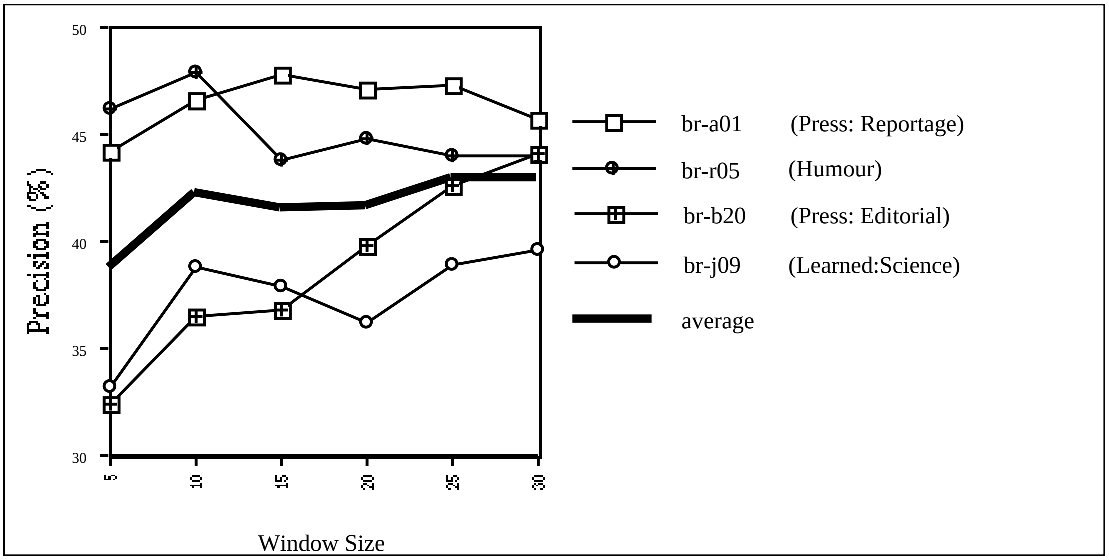


Figure 10: context size and different files

Each text is structured as a list of sentences, lacking any indication of headings, sections, paragraph endings, text changes, etc. This means that the program gathers the context without knowing whether the nouns actually occur in coherent pieces of text. This could account for the fact that in br-r05, composed mainly by short pieces of dialogues, the best results are for window size 10, the average size of pieces of this dialogue. Longer windows will include other pieces of unrelated dialogues that could cause the disambiguation process to go astray.

In addition, SemCor files can be composed of different pieces of unrelated texts without explicit indication. For instance, two of our test files (br-a01 and br-b20) are collections of short journalistic texts. This could explain why the performance of br-a01 decreases for windows of 30 nouns. For most nouns the context window would include nouns from other articles.

The polysemy level could also affect the performance, but in our texts less polysemy does not correlate with better performance. Nevertheless the actual nature of each text is certainly an important factor, difficult to measure, which could account for the different behaviour on its own. For instance, the poor performance on text br-j09 could be explained by its technical nature. Further analysis of the errors, contexts and relations found among the words would be needed to be more conclusive.

In order to give an overall view of the performance, we consider the average behaviour for formulating our conclusions leaving aside these considerations.

5.2.5 partial disambiguation

The disambiguation algorithm has an intermediate outcome between completely disambiguating a word or failing to do so. In some cases the algorithm just manages to discard some senses of the word, but can not choose a single sense. The automatic evaluation program does not take these cases into account, treating them as failures to disambiguate. While the number of words that are not disambiguated decreases for the benefit of completely disambiguated as the window size is bigger, the number of partially disambiguated words stays the same (see Figure 11).

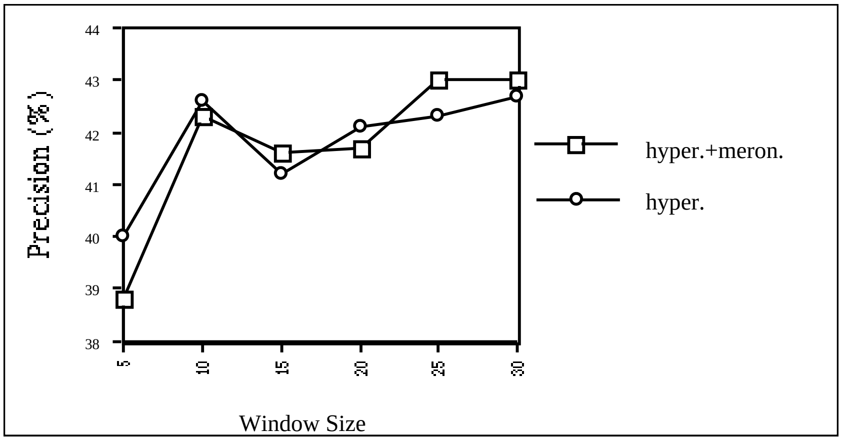


Figure 8: meronymy and hyperonymy

5.2.3 global nhyp is as good as local nhyp.

There was an aspect of the density formula which we could not decide analytically and which we wanted to check experimentally. It refers to the way *nhyp* is calculated (c.f. formula 2). If *nhyp* is computed using formula 2, we call it *local nhyp*, because it has to be computed for every concept of WordNet. Rather than using this *local nhyp*, it would be more desirable, specially for efficiency, if only one global *nhyp* were used for all the concepts. This *global nhyp* can be computed using the whole noun hierarchy. Depending on which *nhyp* is chosen will either be the real number of descendant senses for c (for local *nhyp*) or and estimation based on the global *nhyp*.

To decide whether using local *nhyp* or global *nhyp* affects the performance, we ran parallel experiments using both. The results (see Figure 9) show that there is only a slight difference between them. Therefore, *global nhyp* was used in the experiments.

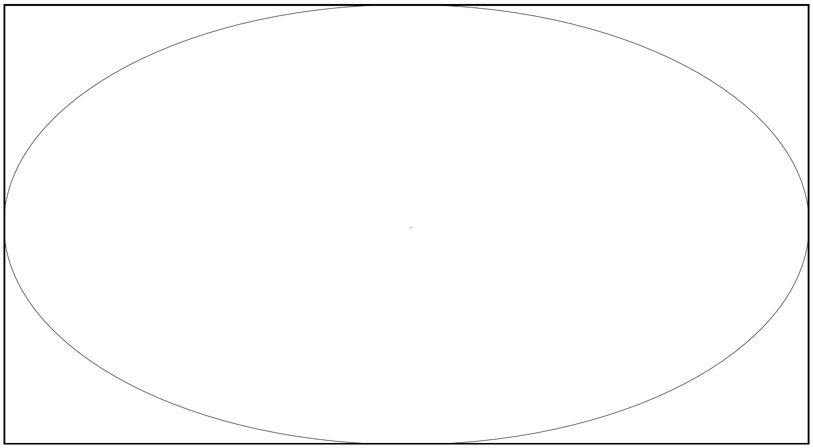


Figure 9: local *nhyp* vs. global *nhyp*

5.2.4 context size: different behaviour for each text

Deciding what context size was better for disambiguating using Conceptual Density is an important issue. One could assume that the more context there is, the better would be the disambiguation results. Our experiments show that each file from SemCor has a different behaviour (see Figure 10). While br-b20 shows clear improvement for bigger window sizes, br-r05 gets a local maximum at a size window of 10 nouns, etc.

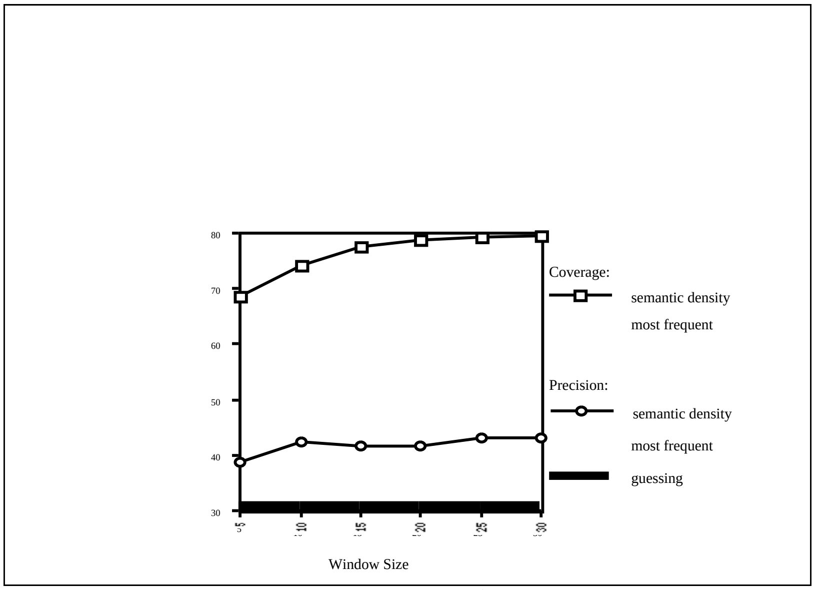


Figure 7: precision and coverage

The figure also shows the guessing baseline, given by selecting senses at random. First, it was calculated analytically using the polysemy counts for the files, which gave 30% of precision. This result was checked experimentally running an algorithm ten times over the files, which confirmed the previous result.

We also compare the performance of our algorithm with that of the most frequent heuristic. The frequency counts for each sense were collected using the rest of SemCor, and then apply the results to the four texts. While the precision is similar to that of our algorithm, the coverage is 8% worse.

All the data for the best window size can be seen in table 2.

%	w=30	Cover.	Prec.	Recall
overall	File	86.2	71.2	61.4
	Sense		64.5	55.5
polysemic	File	79.6	53.9	42.8
	Sense		43	34.2

Table 2: overall data for the best window size

The precision and coverage shown in all preceding plots were relative to the polysemous nouns only. If we also include monosemic nouns precision raises from 43% to 64.5%, and the coverage increases from 79.6% to 86.2%.

5.2.2 meronymy does not improve performance as expected.

One parameter controls whether meronymic relations, in addition to the hypo/hypernymy relation, are taken into account or not. In principle the more relations are taken in account, the better density would capture semantic relatedness and, therefore, the better the expected results. The experiments (see Figure 8) showed that there is not much difference; adding meronymic information does not improve precision, and raises coverage only 3% (approximately). Nevertheless, in the results reported, meronymy and hypernymy were used.

```
<wd>operation</wd><sn>[noun.state.0]</sn><tag>NN</tag>
```

```
<wd>Police_Department</wd><sn>[noun.group.0]</sn><tag>NN</tag>
```

```
<wd>prison_farms</wd><mwd>prison_farm</mwd><msn>[noun.arti-
fact.0]</msn>
    <tag>NN</tag>
```

```
</s>
```

Figure 5: Semcor format

After erasing the irrelevant information we get the following words[6]:

```
jury administration operation Police_Department
                  prison_farm
```

Figure 6: input words

The algorithm then produces a file with sense tags that can be compared automatically with the original file (see figure 5). An automatic program counts sense level matches and file level matches (see Section 5.2.6) for the three classes of results: complete disambiguation, partial disambiguation and failure to disambiguate. For the results shown in Section 5.2, partial disambiguation was considered as failure to disambiguate.

5.2 Results and evaluation

One of the goals of the experiments was to decide among different variants of the Conceptual Density formula. Results are given averaging the results of the four files. Partial disambiguation is treated as failure to disambiguate. Precision[7] is given in terms of polysemous nouns only. Plots are drawn against the size of the context[8] that was taken into account when disambiguating.

5.2.1 evaluation of the results

Figure 7 shows that, overall, coverage of polysemous nouns increases significantly with the window size, without losing precision. Coverage tends to stabilised near 80%, getting little improvement for window sizes bigger than 20.

[6] *Note that in the input texts we already have the knowledge that police department and prison farm are compound nouns, and that the lemma of prison farms is prison farm.*

[7] *Precision is defined as the ratio between correctly disambiguated senses and total number of answered senses. Coverage is given by the ratio between total number of answered senses and total number of senses. Recall is defined as the ratio between correctly disambiguated senses and the total number of senses.*

[8] *context size is given in terms of nouns.*

procedure is repeated. At this point we start afresh with all senses of the words in the window.

Back in the example, the algorithm has disambiguated **operation_3**, **police_department_0**, **jury_1** and **prison_farm_0** (because this word is monosemous in WordNet), but the word *administration* is still ambiguous. The output of the algorithm , thus, will be that the sense for *operation* in this context, i.e. for this window, is **operation_3**. The disambiguation window will move rightwards, and the algorithm will try to disambiguate *Police_Department* taking as context *administration*, *operation*, *prison_farms* and whichever noun is first in the next sentence.

5 The Experiments

5.1 The texts

We selected four texts from SemCor at random: a press report (br-a01), an editorial (br-b20), a scientific text (br-j09) and a humorous article (br-r05). Table 1 shows some statistics for each text

text	words	nouns	nouns in WN	monosemous
br-a01	2079	564	464	149 (32%)
br-b20	2153	453	377	128 (34%)
br-j09	2495	620	586	205 (34%)
br-r05	2407	457	431	120 (27%)
total	9134	2094	1858	602 (32%)

Table 1

An average of 11% of all the nouns in these four texts were not found in WordNet. According to this data, the percentage of monosemous nouns in these texts is bigger (32% average) than the one calculated for the open-class words from the whole SemCor (27.2% according to [Miller et al. 94]). [Sussna 93] presents a similar degree of polysemy for nouns (34% of monosemous nouns), but in a different text collection.

These texts play both the role of input files (without semantic tags) and (tagged) test files. When they are treated as input files, we throw away all non-noun words, only leaving the lemmas of the nouns present in WordNet. The program does not deal with syntactic ambiguity, as the part of speech information is in the input files. Multiple word entries are also available in the input files, as long as they are present in WordNet. Proper nouns have a similar treatment: we only consider those that can be found in WordNet. Figure 5 shows the way the algorithm would input the example sentence in figure 3 after stripping non-noun words:

```
<s>

<wd>jury</wd><sn>[noun.group.0]</sn><tag>NN</tag>

<wd>administration</wd><sn>[noun.act.0]</sn><tag>NN</tag>
```

```
administration_1, governance, government, establishment, brass...

jury_2

   => body

      => people

         => group, grouping
```

Figure 4: partial lattice for the sample sentence

In this example only hypo/hypernym links are shown. The concepts in WordNet are represented as lists of synonyms. Word senses to be disambiguated are shown in bold. Underlined concepts are those selected with highest Conceptual Density. Monosemic nouns have sense number 0.

2) Once the lattice is completed, the program starts the disambiguation loop until there are no words which remains to be disambiguated. For each loop the program computes the Conceptual Density of every concept in the lattice. For instance `<administrative_unit>` has underneath 3 senses to be disambiguated and a subhierarchy size of 96 producing a Conceptual Density of 0.256. Meanwhile, `<body>`, with 2 senses and subhierarchy size of 86, has a Conceptual Density of 0.062.

3) The concept with the highest Conceptual Density (`<administrative_unit>` in our example) is selected.

4) In this step two actions are performed. Firstly the program follows hyponym chains down from the concepts selected in step 3 (`<administrative_unit>`) and the senses of the words found in the bottom are selected as the correct senses (**`operation_3`**, **`police_department_0`** and **`jury_1`** are the senses chosen for *operation*, *Police_Department* and *jury*). All these nouns are considered to be disambiguated, even if more than one sense of a given word are below the concept selected in step 3. Lastly we build the lattice again as in step 1, but only considering the nouns not yet disambiguated. After that, the loop continues in step 2. In the example, the lattice is built for the senses of *administration* and *prison_farms*, but their senses yield non-overlapping lattices (for instance the lattice for **administration_1** would be the same as in figure 4 without **jury_2**), and therefore the loop terminates and we continue in step 5.

5) The program has three possible outcomes for the noun in the middle of the window; one sense has been selected (disambiguated), more than one sense has been selected (partially disambiguated, several senses of the noun are under the same selected concept) or the selection of a sense has been impossible due to the lack of information in the context.

After disambiguating the word in the current window the window moves forward, and the

```
police_department_0
  => local department, department of local government
    => government department
      => department
      jury_1, panel
        => committee, commission
      operation_3, function
        => division
          => administrative unit
            => unit
              => organization
                => social group
                  => people
                    => group, grouping
```

considering the other words in the window as context.

For each window, the program performs the next disambiguation algorithm:

```
(Step 1)tree := compute_tree(words_in_window)
       loop
(Step 2)tree := compute_conceptual_distance(tree)
(Step 3)concept := select_concept_with_highest_weigth(tree)
       if  concept = null then exitloop
(Step 4)tree := mark_disambiguated_senses(tree,concept)
       endloop
(Step 5)output_disambiguation_result(tree)
```

To illustrate the process, consider the following text extracted from SemCor:

> *The jury(2) praised the administration(3) and operation(8) of the Atlanta Police_Department(1), the Fulton_Tax_Commissioner_'s_Office, the Bellwood and Alpharetta prison_farms(1), Grady_Hospital and the Fulton_Health_Department.*

Figure 3: sample sentence from SemCor

The underlined words are nouns represented in WordNet with the number of senses between brackets (those with a 1 are unambiguous nouns). SemCor links multiword terms using underscores. The noun to be disambiguated in our example is *operation.*, and a window size of five will be used.

1) Given the set of nouns constrained by the context window size, our algorithm collects for every noun all its possible senses and hypernyms. All these concepts and connections are placed in a lattice. For each concept in the lattice, the program also stores the set of words that are generalised by the concept.

The following figure shows partially the lattice for the example sentence. Since *Prison_farm* appears in a different hierarchy we do not show it in figure 4:

5.*In fact the algorithm can disambiguate all the nouns in the window in one go, but we consider that the context is most informative for the noun in the center of the window. This and related issues are discussed in Section 6.*

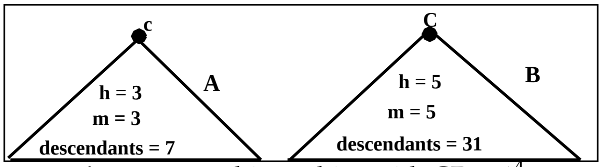


Figure 2: two hierarchies with CD = 1[4].

In order to tune the Conceptual Density formula, we have carried out several experiments adding two parameters, α and β. The α parameter modifies the strength of the exponential *i* because *h* ranges between 1 and 16 (the maximum number of levels in WordNet) while *m* ranges between 1 and the total number of senses in WordNet. Adding a constant β to *nhyp*, we tried to discover the role of the averaged number of hyponyms per concept. Formula 3 shows the resulting formula.

(3)

After a number of runs which were automatically evaluated, the results showed that β does not affect the behaviour of the formula, a strong indication that this formula is not sensitive to constant variations in the number of hyponyms. On the other hand, different values of α affected the performance consistently, yielding the best results in all the experiments where α was 0.20. The formula which was actually used in the experiments, thus, was the following:

(4)

where is the number of descendant senses of the concept *c*.

We have tested the formula in two different ways (see Section 5). The first one involves the manner in which *nhyp* and are calculated. The second arises from the manner in which the hierarchy is constructed: considering only hypo/hypernymy links, or including meronymic links as well.

4 The Disambiguation Algorithm Using Conceptual Density

The algorithm to disambiguate a given noun w in the middle of a window of nouns W roughly proceeds as follows. First, the algorithm represents in a lattice the nouns in the window, its senses and hypernyms (step 1). Then, the program computes the Conceptual Density of each concept in WordNet according to the senses it contains in its subhierarchy (step 2). It selects the concept c with the highest density (step 3) and select the senses below it as the correct senses for the respective words. If a word from W (step 4):

- has a single sense under c, it has already been disambiguated.
- has no a sense, it is still ambiguous
- has more than one sense with highest density, we can eliminate all the other senses of w, but have not yet completely disambiguated w.

It proceeds then to choose the next concept with highest density, and continues to disambiguate words in W. In the end the senses left for w are analysed and the result is output (step 5). This process will be further explained below.

Given a window size, the program moves the window one word at a time from the beginning of the document towards its end, disambiguating the word in the middle of the window[5] and

[4] *From formulas 1 and 2 we have:*

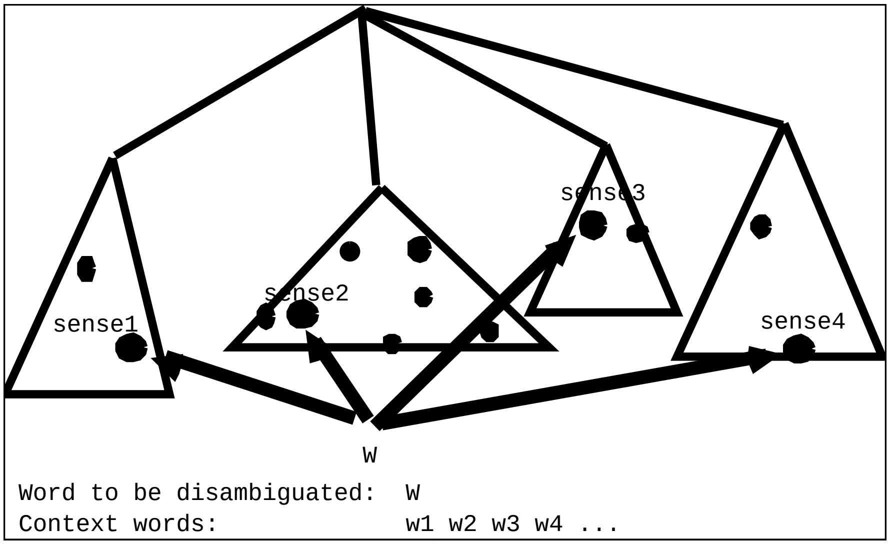


Figure 1: senses of a word in WordNet

The sense of W contained in the subhierarchy with highest Conceptual Density will be chosen as the sense disambiguating W in the given context. In figure 1, sense2 would be chosen.

Given a concept *c*, at the top of a subhierarchy, and given *nhyp* and *h* (mean number of hyponyms per node and height of the subhierarchy, respectively), the Conceptual Density for *c* when its subhierarchy contains a number *m* (marks) of senses of the words to disambiguate is given by the formula below:

(1)

The numerator expresses the expected area for a subhierarchy containing *m* marks (senses of the words to be disambiguated), while the divisor is the actual area, that is, the formula gives the ratio between weighted marks below *c* and the total area of the subhierarchy below *c*. The weight given to the marks tries to express that the height and the number of marks should be proportional.

nhyp is computed for each concept in WordNet in such a way as to satisfy equation 2, which expresses the relation among height, averaged number of hyponyms of each sense and total number of senses in a subhierarchy if it were homogeneous and regular:

(2)

Thus, if we had a concept *c* with a subhierarchy of height 5 and 31 descendants, equation 2 will hold that *nhyp* is 2 for *c*.

Conceptual Density weights the number of senses of the words to be disambiguated so as to make density equal to 1 when the number *m of* senses below *c* is equal to the height of the hierarchy *h*, to make density smaller than 1 if *m* is smaller than *h* and to make density larger than 1 whenever *m* is larger than *h*. The density can be kept constant for different *m*-s provided a certain proportion between the number of marks *m* and the height *h* of the subhierarchy is maintained. Both hierarchies **A** and **B** in figure 2, for instance, have Conceptual Density 1. For the sake of clarity we have assumed uniform hierarchies.

3 Conceptual Density and Word Sense Disambiguation

A measure of the relatedness among concepts can be a valuable predictive knowledge source for several decisions in Natural Language Processing. For example, the relatedness of a certain word-sense to the context allows us to select that sense over the others, and actually disambiguate the word. Relatedness can be measured by a fine-grained conceptual distance [Miller & Teibel 91] among concepts in a hierarchical semantic net such as WordNet. This measure would allow the discovery of the most lexically cohesive set of senses of a given set of words in English.

Several measures of relatedness among words based on cooccurrence in a text have been described; mutual information, t-test, etc. [Church et al. 91], the cosine function in Context Space [Schütze 92], conditional probability [Wilks et al. 93]. [Resnik 93] combines a knowledge based approach involving semantic classes taken from WordNet with cooccurrence data extracted from corpora. Less attention has been paid lately to measures of relatedness based on semantic structured hierarchical nets.

Conceptual distance tries to provide a basis for determining closeness in meaning among words, taking as reference a structured hierarchical net. The conceptual distance between two concepts is defined in [Rada et al. 89] as the length of the shortest path that connects the concepts in a hierarchical semantic net. Besides applying conceptual distance in a medical bibliographic retrieval system and merging several semantic nets, they demonstrate that their measure of conceptual distance is a metric. In a similar approach, [Sussna 93] employs the notion of conceptual distance between network nodes in order to improve precision during document indexing. Following these ideas, [Agirre et al. 94] describes a new conceptual distance formula for automatic spelling correction and [Rigau 94], using this conceptual distance formula, presents a methodology to enrich dictionary senses with semantic tags extracted from WordNet.

The measure of conceptual distance among concepts we are looking for should be sensitive to:

- the length of the shortest path that connects the concepts involved.
- the depth in the hierarchy: concepts in a deeper part of the hierarchy relatively closer than those in a more shallow part.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.

But also:

- the measure should be independent of the number of concepts we are measuring.

We have experimented with several formulas that follow the four criteria presented above. Currently, we are working with a variant of conceptual distance which we call Conceptual Density that compares areas of subhierarchies.

As an example of how Conceptual Density can help to disambiguate a word, in figure 1 the word W has four senses and several context words. Each sense of the words belongs to a subhierachy of WordNet. The dots in the subhierarchies represent the senses of either the word to be disambiguated (W) or the words in the context. Conceptual Density will yield the highest density for the subhierarchy containing more senses of those, relative to the total amount of senses in the subhierarchy.

mantic Concordance or Semcor for short. We also use a public domain lexical knowledge resource, WordNet [Miller 90]. The advantage of this approach is clear; Semcor provides an appropriate environment for testing our procedures in a fully automatic way.

This paper presents a general automatic decision procedure for lexical ambiguity resolution based on a formula of conceptual distance among concepts: Conceptual Density. The procedure needs to know how words are clustered in semantic classes and how semantic classes are hierarchically organised. For this purpose, we have used a broad semantic taxonomy for English, WordNet. We have performed several experiments employing the notion of Conceptual Density among concepts in a structured hierarchical net. Given a piece of text from the Brown Corpus, our system tries to resolve the lexical ambiguity of nouns finding the combination of senses from a set of nouns in context that maximises the total Conceptual Density among senses.

In order to test our algorithms, we have selected at random four texts of SemCor. Our procedure only considers the words in SemCor with a noun part of speech tag. We discarded the nouns not present in WordNet (averaging around 10% of the nouns in all four texts)

Improvement in disambiguation compared with chance is clear and consistent, strongly suggesting that knowledge-based algorithms are competitive with statistically-based approaches, with the advantage of not needing training.

Even if this technique is presented as stand-alone, it is our belief, following the ideas of [McRoy 92] that full-fledged lexical ambiguity resolution should combine several information sources. Conceptual Density might be only one of a number of complementary sources of evidence for evaluating the plausibility of a certain word sense.

In section 2 we present the semantic knowledge sources used by the system. In section 3, we define Conceptual Density. In section 4, we discuss the disambiguation algorithm used in the experiment and in section 5, we explain and evaluate the experiments performed. In section 6, we discuss future directions and, finally, in the last section, we draw some conclusions.

2 WordNet and the Semantic Concordance

Sense is not a well defined concept and often has subtle distinctions in topic, register, dialect, collocation, part of speech, etc. For the purpose of this study, we take as the senses of a word those senses provided by WordNet 1.4 [Miller 90].

WordNet is an on-line lexicon based on psycholinguistic theories. It comprises nouns, verbs, adjectives and adverbs, organised around semantic relations, such as: synonymy and antonymy, hypernymy and hyponymy, meronymy and holonymy. Lexicalised concepts, represented as sets of synonyms called synsets, are the basic elements of WordNet. The senses of a word are represented by synsets, one for each word sense. The version used in this work, WordNet 1.4, contains 83,800 words, 63,300 synsets (word senses) and 87,600 links between concepts.

The nouns of WordNet can be viewed as a tangled hierarchy of hypo/hypernymy relations. Nominal relations include also three kinds of meronymic relations, which can be paraphrased as "member-of", "made-of" and "component-part-of".

SemCor [Miller et al. 93] is a corpus where part of speech and word sense tags (which correspond to WordNet synsets) have been included for all open-class words. SemCor is a subset taken from the Brown Corpus [Francis and Kucera, 67] which comprises approximately 250,000 words from a total of 1 million words. The coverage in WordNet of the senses for open-class words in SemCor reaches 96% according to Miller et al. The tagging was done manually, and the error rate reported is around 10% for polysemous words.

An Experiment on Word Sense Disambiguation of the Brown Corpus using WordNet[1]

Eneko Agirre.[2]*
Departamento de Lenguajes y Sistemas Informáticos. Universidad del Pais Vasco.
p.k. 649, 20080 Donostia. Spain. jibagbee@si.ehu.es

German Rigau.[3]**
Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. g.rigau@lsi.upc.es

Abstract.

This paper presents a method for the resolution of lexical ambiguity and its automatic evaluation over the Brown Corpus. The method relies on the use of the wide-coverage noun taxonomy of WordNet and the notion of conceptual distance among concepts, captured by a Conceptual Density formula developed for this purpose. This fully automatic method requires no hand coding of lexical entries, hand tagging of text or any kind of training process. The results of the experiments have been automatically evaluated against SemCor, the sense-tagged version of the Brown Corpus.

Keywords: Word Sense Disambiguation, Conceptual Distance, WordNet, SemCor.

1 Introduction

Word sense disambiguation is a long-standing problem in computational linguistics. Much of recent work in lexical ambiguity resolution offers the prospect that a disambiguation system might be able to input unrestricted text and tag each word with the most likely sense with fairly reasonable accuracy and efficiency. The main idea is to attempt to use the context of the word to be disambiguated together with information about each of its word senses to solve this problem.

Several interesting experiments have been performed in recent years using pre-existing lexical knowledge resources. [Cowie et al. 92] and [Guthrie et al. 93] describe a method for lexical disambiguation of text using the definitions in the machine-readable version of the LDOCE dictionary as in the method described in [Lesk 86], but using simulated annealing for efficiency reasons. [Yarowsky 92] combines the use of the Grolier encyclopaedia as a training corpus with the categories of the Roget's International Thesaurus to create a statistical model for the word sense disambiguation problem with excellent results. [Gale et al. 93] explains a statistical approach using bilingual parallel corpora. [Wilks et al. 93] perform several interesting statistical disambiguation experiments using cooccurrence data collected from LDOCE. [Sussna 93], [Voorhees 93] and [Richarson et al. 94] define disambiguation programs based in WordNet with the goal of improving precision and coverage during document indexing.

Although each of these techniques looks somewhat promising for disambiguation, either they have been only applied to a small number of words, a few sentences or they are not in a public domain corpus. For this reason we have tried to disambiguate all the nouns from texts in the sense tagged version of the Brown corpora [Francis & Kucera 67], [Miller et al. 93], also called Se-

[1] .* *The research described in this paper was done during a stay in the Computer Research Laboratory, New Mexico State University, Las Cruces, New Mexico.*

[2] .** *Eneko Agirre was supported by a grant from the Basque Government.*

[3] .*** *German Rigau was supported by a grant from the Ministerio de Educación y Ciencia.*

CONCEPTUAL DISTANCE AND AUTOMATIC SPELLING CORRECTION

E. Agirre, X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola

Informatika Fakultatea, p.k. 649. 20080 DONOSTIA
(Basque Country - Spain)
e-mail: jibagbee@si.ehu.es
tel: 34 43 218000

ABSTRACT. Text from different sources usually arrives under imperfect conditions. When an anomalous word is detected automatic word recognisers produce a list of candidates from which only one is correct. A variety of techniques have been devised to discriminate among the possible correction candidates. The project we are involved in tries to exploit linguistic knowledge in Spelling Correction. A preliminary investigation shows syntactic discrimination not to be enough. The gap could be covered by semantic techniques like conceptual distance. Basically, we define conceptual distance between two concepts as the shortest path length in the hierarchies of the lexical knowledge base of IDHS (Intelligent Dictionary Help System). We consider that a correction proposal that is closer to the surrounding words in the sentence is more plausible enabling us to produce a ranking of the proposals. It is our belief that conceptual distance can be also applied to other word recognition areas, such as handwriting recognition or optical character recognition, where a single proposal would also be desirable.

1 INTRODUCTION

Text from different sources usually arrives under imperfect conditions. The medium of transmission conditions the type of automatic word recognition to be used: Optical Character Recognition, Speech Processing or Spelling Correction. When an anomalous input is encountered these recognisers produce a list of candidates from which only one is correct. There are a number of applications e.g. Text-to-Speech Synthesis, that in order to rule out human intervention need automatic correction, that is, the first choice of the correct proposal among the correction candidates.

The task of choosing the appropriate correction proposal is not an easy one. We have to draw knowledge from several sources, as one technique alone would not suffice. In this direction [Kukich, 92] points out, for spelling correction and considering isolated words only, that automatic correction performed by humans scored from %65 to %82. These figures could represent an upper bound for automatic techniques that do not take context into account. To leave %35-%18 of the detected errors uncorrected would be unsatisfactory for the applications mentioned earlier. In order to increase the performance and get an acceptable correction rate, some sort of context modelling, linguistic or other, would be needed.

The project we are involved in tries to exploit linguistic knowledge for automatic spelling correction. This paper focuses on the contribution of lexical-semantic techniques in general, and conceptual distance in particular. Some other work is being carried on the syntactic side.

The idea of conceptual distance captures the intuition that some words are more related or closer than others. We consider that a correction proposal that is closer to the surrounding words in the sentence is more plausible. Thus we can produce a ranking of the proposals.

Basically, we define conceptual distance between two word senses as the shortest path length in the hierarchies of the Dictionary Knowledge Base of IDHS (Intelligent Dictionary Help System [Artola, 93; Agirre et al., 94]), following the ideas of [Rada et al., 87]. The knowledge base of IDHS is a semantic network of frames where each frame represents a word sense from a dictionary. Arcs between frames represent lexical-semantic relations derived from the definitions in a machine readable dictionary.

Next section shows some experimental results that indicate the need of more linguistic knowledge beyond syntax in spelling error correction, followed by an overview of IDHS. After that, two prospective semantic techniques are introduced, from which conceptual distance is explored in depth in the next section. Finally some conclusions are presented.

Originally, the target language was Basque, but later developments in IDHS made us switch to French. For this reason the preliminary collection of data was done for Basque, while the implementation is being run on French texts. The examples in sections 2 and 4 are in Basque, while those in section 5 are in French.

2 ON THE NEED OF SEMANTIC DISCRIMINATION

In order to have some hard data on the convenience and prospective performance of the semantic contribution to automatic error correction, the analysis of a small corpus was performed. The error detection and the list of proposals have been taken from the spelling checker/corrector XUXEN [Aduriz et al, 1993; Agirre et al., 1992]. The texts come from 48 Basque language learners, giving a total of 8290 words. XUXEN generated proposals for 305 spelling errors, producing multiple proposals 182 times (60%).

The syntactic analysis of the texts, as well as the syntactic discrimination of the proposals, was performed by a person simulating an automatic full-fledged and robust parser. The proposals which would lead to grammatical errors where thus removed from the proposal lists. The semantic discrimination was applied only after the syntactic phase was completed.

The results hold that syntax alone could select one single proposal 70% of the cases. This result might be too optimistic, considering that the syntactic analyser was supposed to be complete and robust.

The semantic information faced the cases where syntax could not do the job. Applying by hand the semantic techniques explained below, it managed to solve 63% of the misspellings. It might be that this experiment favoured syntax, leaving semantics the tough cases. Anyway, the performance of both is similar, and the experiment indicates that their combination is desirable in order to get better results, up to 90% in this particular experiment. These results are tentative, awaiting confirmation of implemented systems with realistic syntactic and semantic coverage.

XUXEN: 305 errors with proposals		
1 prop.	123	40.3%
n prop.	182	59.7%
syntactic discrimination on 182 errors		
success	128	70.3%
fail	54	29.7%
semantic discrimination on 54 errors		
success	34	62.9%
2/3	11	20.3%
fail	9	16.8%

3 IDHS

IDHS (Intelligent Dictionary Help System) provides the base for semantic correction. It provides both a representation language suited to explore the techniques presented in the following section, and also the semantic knowledge itself.

IDHS was conceived as a monolingual (explanatory) dictionary system for human use [Artola & Evrard, 92; Artola, 93]. The system provides various access possibilities to the data, allowing to deduce implicit knowledge from the explicit dictionary information. The system has been implemented on a symbolic architecture machine using KEE knowledge engineering environment.

The starting point of IDHS is a Dictionary Database (DDB) built from an ordinary French dictionary. Meaning definitions have been analysed using linguistic information from the DDB itself and interpreted to be structured as a Dictionary Knowledge Base (DKB). As a result of the parsing different lexical-semantic relations between word senses are established by means of semantic rules (attached to the patterns); rules are used for the initial construction of the DKB.

The interconceptual lexical-semantic relations detected from the analysis of the source dictionary are classified into paradigmatic and syntagmatic. Among the paradigmatic relations, the following have been found: synonymy and antonymy, taxonomic relations as hypernymy/hyponymy —obtained from definitions of type "genus et differentia"— and taxonymy itself (expressed by means of specific relators such as sort-of and kind-of), meronymy, and others. Whereas among the syntagmatic relations we can find case relations (e.g. agent, object, goal, etc.), relations derived from the specific lexicographic metalanguage (e.g. quality-of, act-of, property), and others.

The knowledge representation scheme chosen for the DKB of IDHS is composed of three elements, each of them structured as a different knowledge base. One of this components, KB-THESAURUS, implements the dictionary as a semantic network of frames, where each frame represents a one-word concept (word sense) or a phrasal concept. Phrasal concepts represent phrase structures associated to the occurrence of concepts in meaning definitions. Frames are interrelated by slots representing lexical-semantic relations. Other slots contain phrasal, meta-linguistic, and general information.

In the following section we tackle spelling correction from the point of view of semantics and IDHS.

4 SEMANTIC DISCRIMINATION

As we already mentioned, this work focuses primarily on the contribution of semantics, and more precisely in the use of lexical-semantic information. We have considered the use of the following:

Selectional Restrictions

Selectional restrictions indicate semantic constraints that the arguments of verbs, adjectives or nouns have to fulfil. For example:

```
jan      => verb[agent: animate,
                 object: edible]
ilegorri => adj.[argument: person]
anaia    => noun[argument: person]
```

These can be read as 'the verb jan (eat) takes as agent an animate entity and as object and edible entity', 'the argument of ilegorri (blonde) has to be a person', etc.

The contribution of selectional restrictions will be illustrated by the following example from the Basque corpus. Had someone typed lehio in Basque we would get the proposals below[1]:

```
lehio: lehia, lesio, leiho
```

If the misspelling occurs in the following sentence, and assuming a sample selectional restriction for apurtu (to break),

```
"lehio bat apurtu dut"[2]
apurtu =>  [agent: animal,
            object: physical-object]
```

we would be able to discard competition and injury, and select the only proposal that fulfils the restriction of being a physical object, leiho (window).

Conceptual Distance

The idea of conceptual distance tries to capture the intuition that some words are closer or more related than others. Therefore we can consider devising a metric that would give results similar to the following[3]:

```
dist(itsasontzi,kapitain) = "short"
dist(itsasontzi,teklatu) = "long"
```

The idea is that we prefer proposals that are related or conceptually close to the other words in the sentence, rather than unrelated or distant proposals. This approach has multiple variants, depending on whether we take all the words in the sentence, or we only take the measurements with some relevant words in the sentence.

Let us consider the following example[4]:

```
uzaina: zaina, usaina, uhaina
"ukenduaren uzainak erlea aldendu zuen"
```

We can compare the distance of the proposals with the other words in the

[1] The proposals mean respectively competition, injury, window.

[2] Meaning *I broke a <lehio>*. All the basque examples and proposals in the paper are taken from a small corpus and the correction proposals are all from Xuxen

[3] The words mean respectively *ship, captain, keyboard.*

[4] The proposals mean, respectively, *vein, smell, wave.* The sentence means *the <uzaina> of the ointment kept away the bee.*

sentence. The result would be that `usaina` (smell) holds the minimum total distance, and therefore would be preferred as the correct proposal. This technique will be further explained below.

5 CONCEPTUAL DISTANCE AND SPELLING CORRECTION

Mainstream approaches to conceptual distance rely on structured inheritance nets or similar kinds of knowledge bases. For instance, [Rada et al., 89] defines conceptual distance in terms of the length of the shortest path of IS-A links between the word senses of the Mesh semantic net. Besides applying distance in a medical bibliographic retrieval system, they also try to use it as a tool for merging semantic nets.

In a similar approach, [Sussna, 93] assigns a weight to each link in the Wordnet semantic network and calculates the distance between two word senses as the total weight of the path with minimum weight. The weights try to capture additional data, e.g. tfor the same path length, word senses lower in the hierarchy seem to be conceptually closer.

These two approaches take into consideration that words have multiple senses. In fact [Sussna, 93] devises his measure with the purpose of sense-disambiguating a text for indexing and text retrieval.

The knowledge representation of IDHS provides support for the experimentation of several distance measures, allowing us to select the most suitable for proposal discrimination. Previous works on conceptual distance rely mainly on hierarchical relations (hypernymy, taxonymy, meronymy), but distance measures could also profit from the other semantic relations in IDHS. [Rada et al., 89] point out that the proliferation of semantic relations makes distance unreliable. Such systems (e.g. [Collins et Loftus, 75]) have to provide a complex weighting mechanism to balance the heterogeneous nature of the relations. In order to avoid that, it would be desirable to use certain semantic relation only when appropriate, that is, when it makes sense in the given context. This idea will be developed below, while considering the issues related to the application of conceptual distance to correction.

Path-Finding Algorithms

In the heart of the distance algorithm there is a path-finding algorithm. Given two word senses in IDHS, the algorithm would find the shortest path(s) of lexical-semantic links between both. In order to be able to test different correction strategies the following algorithms have been implemented:

`h-path(n1,n2)`: finds the path following hierarchical links only: hypernym, part-of, component-of, element-of, sort-of and their respective inverse relations.

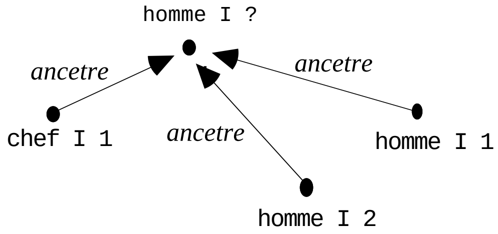
`s-path(n1,n2,r1,...,rn)`: finds a path that has to contain at least one non-hierarchical (semantic) link from the set {r1,...,rn}, alongside the previously mentioned hierarchical links.

`s*-path(n1,n2)`: finds a path that may contain any non-hierarchical (semantic) relation, alongside the hierarchical links.

The first algorithm, `h-path`, constraints the search to hierarchical relations only. It is considered the most reliable for conceptual distance, but it imposes several limitations. The two word senses need to be in the same hierarchy, which implies that `h-path` will never find a path across different parts of speech. For the same reason, it needs very comprehensive hierarchies, which are difficult to create or acquire. Other semantic relations could alleviate this, relating concepts across hierarchies.

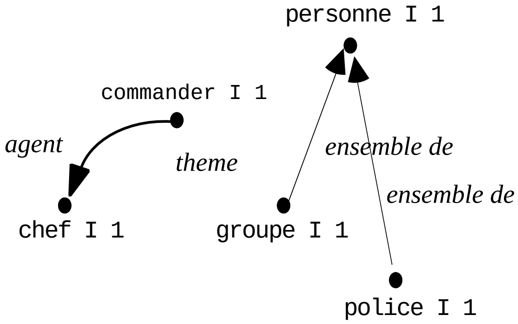
The use of unconstrained semantic relations as in `s*-path`, though, can produce nonsense paths that have to be neutralised when calculating the actual distance figures. It also has heavy efficiency burdens, which can be reduced constraining the set of acceptable relations. If the set of relations is constrained according to semantic criteria, the paths will be semantically coherent. The set of acceptable relations for a certain pair of word senses could be deduced from context, or in some cases, from the part of speech of the word senses. For instance, IDHS admits two relations for a noun that have an adjective as value: property and quality-of. In that case `s-path` will return a path that relates both noun and adjective via property, quality-of and the hierarchical relations.

Some examples of the algorithms follow:



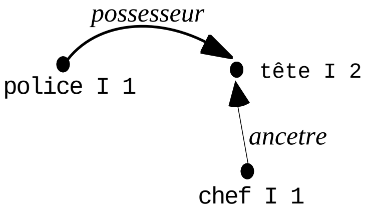
h-path(chefI1, hommeI1) =
chefI1 *ancetre* hommeI? *descendant* hommeI1

The path found by `h-path` between the first word sense of *boss* and the first sense of *man* means: *bossI1 is an ancestor*[5] *of manI?* (a non-disambiguated sense that includes all other senses of man), *which has as descendant manI1.*



s*-path(chefI1, policeI1) =
chefI1 *agent+inv* commanderI1 *theme* groupeI1 *ensemble de* personneI1 *element de* policeI1

The path found by `s*-path` between the first word sense of *boss* and the first sense of *police* means: *bossI1 is an agent of to-commandI1 which has as object groupI1, which is a set-of personI1 which is an element-of policeI1.*



s-path(chefI1,policeI1,possesseur+inv) =
chefI1 *ancetre* têteI2 *poss.+inv* policeI1

The path found by `s*-path` between the first word sense of *boss* and the first sense of *police* means: *bossI1 is a descendant of headI2 (as in head of department), which is "owned" by policeI1.*

The general search of a path between two nodes has exponential complexity, in the order of $O(c^n)$, where `c` is the average of the number of links per word sense, and `n` is the length of the path. In order to keep it under control, the length of the path has to be limited beforehand. This limit can be interpreted as the point after which we consider the two nodes to be unrelated or "very" far. Accordingly, this limit should be "tuned" having in consideration both efficiency and conceptual suitability.

The complexity of the three algorithms grows from the first to the last. While `h-path` deals with five hierarchical relations ($c \leq 5$) and `s-path` is devised to also take into account a small set of relations of the same kind (one to four extra relations, $c \leq 9$), `s*-path` has to provide for the whole set of relations (ranging from 10 to 40 depending on the part of speech of the word sense).

Conceptual Distance

The path(s) between two word senses is(are) the base for conceptual distance. But other facts have to be also considered. The empirical results of [Sussna, 93] show that, as already mentioned at the beginning of this section, the length of the path and the specificity of the word senses in the path (measured by the depth in the hierarchy) are the important parameters that affect the distance measure he proposes. The second parameter tries to capture the fact that specific word senses are considered closer than more general ones.

Our conceptual distance reflects those parameters in the following formula:

$$\mathbf{distance}(\boldsymbol{ws_1}, \boldsymbol{ws_n}) = \sum_{i=1}^{n} \mathbf{1/depth}(\boldsymbol{ws_i})$$

where $< ws_1 K\ ws_i K\ ws_n >$ is the path from ws_1 to ws_n, and $\mathbf{depth}(ws_i)$ is the depth of ws_i in the taxonomy.

Other parameters that could help tuning the measure have not been considered yet. One parameter, for example, could involve giving different weights to each relation, in a way similar to the "criteriality tags" used by [Quillian, 68]. The inclusion of these parameters in the above formula depends greatly on empirical results, which have not yet been gathered.

Correction

As mentioned in section 4, we perform correction choosing the proposal that is more

[5] Ancestor includes the concepts in the transitive closure of hypernymy. Descendant includes the concepts in the transitive closure of hyponymy.

related or conceptually closer to the other words in the sentence, and leaving aside unrelated or distant proposals. The relatedness of a given proposal with the surrounding sentence can be measured using a variety of strategies.

g-correction (generalised). Distance as defined above is measured between word senses. Consequently all the senses in the dictionary for the words in the sentence and the proposals have to be considered. This means that inappropriate senses could bias the corrector to choose an incorrect proposal. In order to rule out, or at least try to neutralise, these spurious readings, and at the same time choose the correct proposal, the following technique can be used: the preferred senses and proposals will be the ones that give minimal pairwise conceptual distance.

Thus, if we have a sentence of length N $<w_1, w_2, ...w_n>$ with M spelling errors $\{e_1=w_i...e_m=w_j\}$, and a list of proposals for each error $P(e_i) = <p_{i1},...p_{iL}>$, we need to consider the senses of all non-error words and the proposals. For each possible combination of senses (mixing both non-error words and proposals), the winning combination will be the one with the minimal total of pairwise distances. This winning combination will give both the preferred proposals and word senses.

In figure 1, it can easily be seen that for long sentences with highly ambiguous words and many correction proposals, the number of combinations and pairwise distance computations grows enormously.

c-correction (constrained). If we want to limit both the number of combinations and the pairwise distance computations, we can focus on doing proposal discrimination only. We are not trying to sense-disambiguate now, and will thus consider of equal value incorrect word senses and appropriate ones.

For each proposal we will only compute the distances of its corresponding word senses with each word sense of the non-error words in the sentence (cf. fig. 2). The proposal that gets the minimum total distance wins.

Sentence: le cheé de la police reunit vingt hommes sur la place du village.
Error: cheé Proposals: chef cher chez chié chieé chéri chic

Word Senses in IDHS:

Sentence: police I 1, police I 2,
reunir I 1, reunir I 2, reunir I 3, reunir I 4, reunir I 5
homme I 1, homme I 2, homme I 3, homme I 4, homme I ?
place I 1, place I 2, place I 3, place I 4, place I 5, place I ?
village I 1

Proposals: chef I 1, cher I 1, cher I 2, chéri I 1, chic I 1

Combinations:

C1) police I **1**, reunir I **1**, homme I 1, place I 1, village I 1, chef I 1
C2) police I **2**, reunir I **1**, homme I 1, place I 1, village I 1, chef I 1
...

Number of combinations: 2x5x5x6x1x5 = 1.500

Distance on C1:

```
dist(police I 1, reunir I 1) ... dist(police I 1, chef I 1)     n=5
dist(reunir I 1, place I 1)  ... dist(reunir I 1, chef I 1)     n=4
...
dist(village I 1, chef I 1)                                     n=1
```

Number of distance calls:

```
[total]   1500 x (5+4+3+2+1) = 1500 x 15 = 22.500
[distinct pairs]                             239
```

fig. 1. Combinations in **g-correction**.[6]

[6] The sentence means "the *cheé* of the police gathered twenty men in the square of the village". The proposals for cheé are: boss, expensive, ´home of´, dear and stylishness..

```
Combinations:
  chef I 1  police I 1, police I 2,
            reunir I 1, reunir I 2, reunir I 3, reunir I 4, reunir I 5
            homme I 1, homme I 2, homme I 3, homme I 4, homme I ?
            place I 1, place I 2, place I 3, place I 4, place I 5, place I ?
            village I 1
  ...
  chic I 1  police I 1, police I 2,
            reunir I 1, reunir I 2, reunir I 3, reunir I 4, reunir I 5
            homme I 1, homme I 2, homme I 3, homme I 4, homme I ?
            place I 1, place I 2, place I 3, place I 4, place I 5, place I ?
            village I 1

      Number of combinations:  5

Distance:
  C1)       dist(chef I 1, police I 1) ... dist(chef I 1, village I 1)
            ...
            dist(chic I 1, police I 1) ... dist(chic I 1, village I 1)

      Number of distance calls:
        [total]    5x(2+5+5+6+1)= 95
```

fig. 2. Combinations in **c-correction**.

Although the wrong word sense may contribute to credit incorrect proposals, the greater number of related true senses will add up and eventually the correct proposals will be chosen.

s-correction (*"semantic"*). We have already introduced two path-finding algorithms (s-path and s*-path) that traverse non-hierarchical semantic relations. The semantic clues in the sentence can be used to inform s-path about the relations that can be expected in the path between the two word senses. Figure 3 illustrates a simplified example of the semantic relations in the sentence from figure 1. The preposition *de* can be interpreted as meaning owner, location etc. For the example below, calling s-path with the corresponding word senses will find a path. We already saw an example when examining path-finding.

This kind of semantic interpretation does not require as heavy a linguistic machinery as it might seem. Triples like those of the example are readily obtained by semantic information extraction systems from corpora [Velardi et al., 91].

6 CONCLUSIONS AND FURTHER WORK

We have outlined the application of a specific semantic technique, conceptual distance, in automatic spelling correction.

```
Semantic relations:

   from the verb:
   (reunit agent cheé)
   ...
   from the preposition de:
   (cheé possesseur+inv police)
   (cheé location police)
   ...
Combinations & Distance:
   reunir I 1 chef I 1...chic I 1
   ...
   reunir I 5 chef I 1...chic I 1

   chef I 1...chic I 1 police I 1
   chef I 1...chic I 1 police I 2
   ...

   Number of combinations:   5+2+2=9
   Number of dist. calls:       9x5=45
```

fig. 3. Combinations in **s-correction**.

In previous implementations of conceptual distance, only h-path style algorithms have been used. These algorithms need comprehensive hierarchies, which are difficult to construct. Other semantic relations. i.e. non-hierarchical relations, can serve to relate word senses even if they do not share the same hierarchy, and specially in the case of two word senses from different grammatical categories. These extra semantic relations could be exploited by conceptual distance using s*-path and s-path. Selectional restrictions are also an alternative in this kind of situations.

s*-path has coherence and efficiency problems which are alleviated in s-path. But in order to use s-path properly, semantic information from the context of the error has to be obtained. This semantic analysis and the tuning of the specific relations needed in a certain context are the work we are focusing on now.

In a further step, we are also planning to develop a more efficient application-oriented representation of the semantic knowledge. For that purpose, we will try to identify and map the relevant subset of the representation of IDHS.

Other important issue is the application of the different correction strategies to real data, where their performance should be effectively contrasted. In this sense, IDHS, because of the rich variety of semantic relations extracted from the dictionary, is very well suited as a platform for extensive testing of the issues above.

It is our believe that the correction techniques explored in this paper, although originally designed for spelling correction, are not dependent of the error source. As long as they are applied on linguistic input they could be used in other word recognition areas where automatic correction, i.e. single correction proposals, would be desirable.

ACKNOWLEDGEMENTS

Eneko Agirre was enjoying a grant from the Basque Government during the present work.

REFERENCES

Aduriz, I., Agirre, E., Alegria, I., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Maritxalar, M., Sarasola, K. and Urkia, M. A Morphological Analysis Based Method for Spelling Correction, in *Proceedings of the E.A.C.L.*, Utrecht, The Netherlands. 1.993

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. and Urkia, M. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology in *Proceedings of the third conference on Applied Natural Language Processing*, Trento, Italy. 1992.

Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. Aproximación funcional a DIAC: Diccionario inteligente de ayuda a la comprensión, *Proc. SEPLN*, 11, 127-138. 1991.

Artola, X. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception d'un système intelligent d'aide dictionnariale (SIAD). PhD thesis. UPV-EHU. 1993.

Artola X., Evrard F. Dictionnaire intelligent d'aide à la compréhension, *Actas IV Congreso Int. EURALEX´90* (Benalmádena), 45-57. Barcelona: Biblograph, 1992.

Collins A.M., Loftus E.F. "A spreading activation theory of semantic processing", *Psych. Rev.*, vol. 82, no. 9, Sept. 1975

Kukich K. Techniques for Automatically Correcting Words in Text, in *ACM Computing sureys*, vol. 24, no. 4. December 1992.

Quillian, M.R. Semantic Memory in M. Minsky ed., p. 227-270, *Semantic Information Processing*. Cambridge (Mass.): MIT Press, 1968.

Rada, R., Mili, H., Bicknell, E. and Blettner, M. Development and Applicarion od a Metric on Semantic Nets, in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, 17-30. 1989.

Sussna, M. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, in *Proceedings of the Second International Conference on Information and Knowledge Management*, Airlington, Virginia USA. 1993.

Velardi P., Fasolo M. and Pazienza M.T. How to encode Semantic Knowledge: a Method for Meaning Representation and Computer-Aided Acquisition, *Computational Linguistics* 17, 2. 1991.

Lexical-semantic information and automatic correction of spelling errors.

E. Agirre, X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola

Informatika Fakultatea, p.k. 649. 20080 DONOSTIA
(Basque Country - Spain)
e-mail: jibagbee@si.ehu.es
tel: 34 43 218000

1. INTRODUCTION.

This study focuses on the use of lexical-semantic information for the automatic discrimination of the proposals generated by a spelling corrector. Current spelling checkers only detect non-word errors, e.g. *sgip*, *shap* instead of *ship*, but would not notice *sip* as a misspelling of *ship*. Moreover, they hand out a list of correction proposals, leaving to the user the decision of which one was the intended word, for instance[1]:

```
araso*:   eraso, arazo, arasa, arbaso
```

In general, it is not possible to guess which one is the correct proposal in isolation, we need to examine the context2:

```
"araso hau konpontzeko eskatu dut."
```

Confronted with this sentence, a Basque speaker would choose 'arazo' (*problem*) as the correct word. A system able to take this decision should include at least syntactic and also semantic information. In the example above, for instance, syntax can not eliminate any proposal, being all from the same syntactic category. Semantic information, on the contrary, strongly indicates that what you *solve* has to be an 'arazo' (*problem*), rejecting the other proposals.

This paper presents firstly an overview of some prospective techniques. In the third section the results of a study in a small corpus are also commented. Next, the way in which IDHS, Intelligent Dictionary Help System [Arregi et al., 1993], can be applied is explored. Finally some conclusions and proposals for future work are suggested

[1] For the misspelled araso, the spelling corrector for Basque Xuxen gives a list of proposals which mean respectively *attack*, *problem*, *shelf* and *ancestor*. All the examples and proposals in the paper are taken from a small corpus and the correction proposals from Xuxen [Aduriz et al., 1993] [Agirre et al., 1992].

[2] The sentence means: *I asked to solve this <araso>*.

2. LEXICAL SEMANTIC TECHNIQUES.

As we already mentioned, this work focuses primarily on the contribution of semantics, and more precisely in the use of lexical-semantic information. We are considering the use of the following:

1) selectional restrictions

Selectional restrictions indicate semantic constraints that the arguments of verbs, adjectives or nouns have to fulfil. For example:

```
eat      => [agent: animate, object: edible]
blonde   => [argument: person]
brother  => [argument: person]
```

These can be read as 'the verb *eat* takes as agent an animate entity and as object and edible entity', 'the argument of *blonde* has to be a person', etc.

The contribution of selectional restrictions will be illustrated by the following example. Had someone typed `lehio` in Basque we would get the proposals below[3]:

```
lehio:    lehia, lesio, leiho
```

If the misspelling occurs in the following sentence, and assuming a sample selectional restriction for `apurtu` (*to break*),

```
"lehio bat apurtu dut"
```
[4]

```
apurtu =>   [agent:animal,
             object: physical-object]
```

we would be able to discard competition and injury, and select the only proposal that fulfils the restriction of being a physical object, `leiho` (*window*).

2) lexical-conceptual distance

The idea of lexical-conceptual distance tries to capture the intuition that some words are closer or more related than others. Therefore we can consider devising a metric that would give results similar to the following:

[3] The proposals mean respectively *competition, injury, window.*

[4] Meaning *I broke a <lehio>*.

```
distance(ship, captain) = "short"
distance(ship, keyboard) = "long"
```

The idea is that we prefer proposals that are related or conceptually close to the other words in the sentence, rather than unrelated or distant proposals. In order to only take the measurements with the relevant words in the sentence, it would be desirable that a syntactic analysis had been performed.

Let us consider the following example:[5]

```
uzaina:   zaina, usaina, uhaina
```

"ukenduaren <u>uzainak</u> erlea aldendu zuen"

We choose to compare the distance of the proposals (which are the subjects of the sentence) with their complement `ukendu` (ointment) and the direct object `erle` (bee). The result would be that `usaina` (*smell*) holds the minimum total distance, and therefore would be preferred as the correct proposal.

```
total = dist(ukendu,X) + dist(erle,X)
```

3 ANALYSIS OF THE ERRORS IN A SMALL CORPUS OF BASQUE

In order to have some hard data on the convenience and prospective performance of the semantic contribution to automatic error correction, the analysis of a small corpus was performed. The error detection and the list of proposals have been taken from the spelling checker/corrector XUXEN. The texts come from Basque language learners, giving a total of 8000 words. From the nearly 500 spelling errors XUXEN detected, 182 errors involved multiple proposals.

The syntactic analysis, as well as the syntactic discrimination of the proposals was performed by a person simulating an automatic parser. The semantic discrimination was applied only to the proposals deemed correct by the syntactic phase.

The results hold that syntax alone could select one single proposal 70% of the cases. This result might be too optimistic, considering that the syntactic analyzer was supposed to be complete and robust. The semantic information was faced with the cases where syntax could not do the job, and managed to solve 63% of the misspellings. The performance of both is similar, and the

[5] The proposals mean, respectively, *vein, smell, wave*. The sentence means *the <uzaina> of the ointment kept away the bee.*

experiment indicate that their combination is desirable in order to get better results, up to 90% in this particular experiment.

4 IDHS AND THE ACQUISITION OF THE REQUIRED LEXICAL-SEMANTIC INFORMATION

One of the motivations of this work is to take profit from the relations and deductive power available in IDHS, which is constructed from conventional dictionaries. Each kind of semantic information is studied in turn:

1) Selectional Restrictions

IDHS does not provide information on selectional restrictions explicitly. It would be desirable to acquire selectional restrictions automatically, and there is some work done in this direction: acquisition from corpora [Velardi et al., 89] [Velardi et al., 91] [Grishman and Sterling, 92] and from codes already provided in machine readable dictionaries for English [Boguraev and Briscoe, 87]. There are not many publications though on the acquisition of selectional restrictions from dictionary definitions.

IDHS was constructed automatically parsing dictionary definitions, and a careful analysis of the information contained in the definitions could give clues to the processing of their representation in IDHS and the automatic acquisition of selectional restrictions. A similar approach proved successful for the acquisition of the aktionsart of English verbs [Alonge, 91]. This process could also profit from the relations already inferred in IDHS, such as synonymy, taxonomy, meronymy, etc. For instance, there is evidence that the selectional restriction information of verbs is specialized down the taxonomy [Calzolari, 90]. Finally the selectional restriction information can be integrated in the representation of IDHS.

2) Lexical-Conceptual distance

Some approaches to distance rely on semantic nets or similar kinds of Knowledge Bases. [Rada et al., 89] define conceptual distance on terms of the length of the shortest path of IS-A links between the concepts. [Sussna, 93] assigns a weight to each link and calculates the distance between two concepts as the weight of the path with minimum weight. The weights try to capture additional data. For instance, for the same path length, concepts lower in the hierarchy seem to be conceptually closer. One further approach [Resnik, 93] combines both corpus-based information-theoretic measures and the taxonomy (implemented as IS-A links) of a semantic net, defining conceptual distance, or conceptual similarity, as a function of the probability of concepts in the training corpus.

All these three approaches take into consideration that words have multiple senses. For instance, [Sussna, 93] devises his measure with the purpose of sense-disambiguating a text for indexing and text retrieval.

The knowledge representation of IDHS provides support for the experimentation of several distance measures, in order to select the most suitable for proposal discrimination. Distance measures could also profit from the other semantic relations in IDHS, as previous works rely mainly on IS-A links. [Rada et al., 89] point out that other relations could be useful, and that further work should be done in this direction. IDHS relates the concepts with a rich variety of semantic relations, such as taxonomy, meronimy or non-hierarchical relations like *theme-of*, *agent-of*, *purpose-of*, *antonymy*, etc. which should be explored.

The thesaurus of IDHS already provides a function that finds relationships between pairs of concepts, called DRAP. The result of this function is a path of concepts in the thesaurus labelled with semantic relations.

The kind of relations found by DRAP are illustrated by the following examples for french:

```
;;; Which is the relation between "couteau I 1" (knife) and
;;; "trancher I ?" (to cut a slice) ?

(drap '|couteau I 1| '|trancher I ?|)
➔    ((AND (|couteau I 1| OBJECTIF |couper I 1|)
           (|couper I 1| SYNONYMES |trancher I ?|)))[6]

;;; Which is the relation between "gazeux I 1" (gaseous) and
;;; "liquide I ?" (liquid) ?

(drap '|gazeux I 1| '|liquide I ?|)
➔    ((AND (|gazeux I 1| CARACTERISTIQUE+INV |vapeur I 2|)
           (|vapeur I 2| CARACTERISTIQUE |liquide I ?|)))[7]
```

[6] Roughly paraphrased as "the purpose of couteau is couper (to cut) which is a synonym of trancher".

[7] Roughly paraphrased as "gazeux is a feature of vapeur (vapour) which has as feature liquide".

```
;;; Which is the relation between "quart I 3" (a beaker of 1/4
;;; l. of capacity) and "vin I 1" (wine)?

(drap '|quart I 3| '|vin I 1|)

➔    ((AND (|quart I 3| OBJECTIF |boire I ?|)
            (|boire I ?| THEME |boisson I 1|)
            (|boisson I 1| HYPONYME |vin I 1|)))[8]
```

5 CONCLUSIONS

The analysis of the corpus confirms that semantic discrimination of proposals is necessary if automatic error correction based in linguistic knowledge is to be obtained, as syntactic discrimination could only succeed maximun 70% of the times, given that all the sentences in the text were completely analyzed.

Both semantic techniques, selectional restriction and semantic distance, can profit from IDHS, which offers a good platform for the acquisition of the former, and the possibility to explore different algorithms for the later.

It has to be noted that a system with the ability to correct automatically spelling errors based on linguistic knowledge, can be also applied to perform automatic error correction in other fields where language is the support of the data, e.g. optical character recognition, text-to-speech systems and speech recognition.

Bibliography.

Aduriz, I., Agirre, E., Alegria, I., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Maritxalar, M., Sarasola, K. and Urkia, M. A Morphological Analysis Based Method for Spelling Correction, in Proceedings of the E.A.C.L., Utrecht, The Netherlands. 1.993

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. and Urkia, M. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology in *Proceedings of the third conference on Applied Natural Language Processing*, Trento, Italy. 1992.

Alonge, A. Extraction of information on aktionsart from verb definitions in machine readable dictionaries, in Avignon 91, vol. 8. 1991.

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, M., Evrard, F and Sarasola, K. IDHS, MLDS: Towards dictionary help systems for human users. *This volume.* 1993.'

Artola, X. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception d'un système intelligent d'aide dictionnariale (SIAD). PhD thesis. UPV-EHU. 1993.

Boguraev, B. and Briscoe, T. Large Lexicons For Natural Language Processing: Utilising The Grammar Coding System of LDOCE, *Computational Linguistics* 13, 3-4, 203-218. 1987.

[8] Roughly paraphrased as "the purpose of `quart` is to `boire` (to drink); the theme of `boire` is `boisson` (drink as a noun) and `vin` is a kind of `boisson`".

Calzolari, N. Structure and access in an automated lexicon and related issuess, in *Linguistica Computazionale Vol VI: Computational Lexicology and Lexicography*, 139-161. 1990.

Grishman R. and Sterling J. Acquisition of selectional patterns, in *Proceedings of COLING-92* (Nantes), 658-664. 1992.

Rada, R., Mili, H., Bicknell, E. and Blettner, M. Development and Applicarion od a Metric on Semantic Nets, in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, 17-30. 1989.

Resnik, P. Semantic Classes and Syntactic Ambiguity, in *Proceedings of the ARPA Workshop on Human Language Technology*. Princeton. 1993.

Sussna, M. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, in *Proceedings of the Second International Conference on Information and Knowledge Management*, Airlington, Virginia USA. 1993.

Velardi P., Fasolo M. and Pazienza M.T. How to encode Semantic Knowledge: a Method for Meaning Representation and Computer-Aided Acquisition, *Computational Linguistics* 17, 2. 1991.

Velardi P. and Pazienza M. T. Computer Aided Interpretation of Lexical Coocurrences in U. Zernik ed.*Proceedings of the 1rst International Lexical Acquisition Workshop* (Detroit), 185-192. 1989.

Towards a single proposal in spelling correction

Eneko Agirre, Koldo Gojenola, Kepa Sarasola
Dept. of Computer Languages and Systems
University of the Basque Country, 649 P. K.,
E-20080 Donostia, Basque Country
eneko@si.ehu.es

Atro Voutilainen
Department of General Linguistics
University of Helsinki, P.O. Box 4
FIN-00014 Helsinki, Finland
avoutila@ling.helsinki.fi

Abstract

The study presented here relies on the integrated use of different kinds of knowledge in order to improve first-guess accuracy in non-word context-sensitive correction for general unrestricted texts. State of the art spelling correction systems, e.g. *ispell*, apart from detecting spelling errors, also assist the user by offering a set of candidate corrections that are close to the misspelled word. Based on the correction proposals of *ispell*, we built several guessers, which were combined in different ways. Firstly, we evaluated all possibilities and selected the best ones in a corpus with artificially generated typing errors. Secondly, the best combinations were tested on texts with genuine spelling errors. The results for the latter suggest that we can expect automatic non-word correction for *all* the errors in a free running text with 80% precision and a single proposal 98% of the times (1.02 proposals on average).

Introduction

The problem of devising algorithms and techniques for automatically correcting words in text remains a research challenge. Existing spelling correction techniques are limited in their scope and accuracy. Apart from detecting spelling errors, many programs assist users by offering a set of candidate corrections that are close to the misspelled word. This is true for most commercial word-processors as well as the Unix-based spelling-corrector *ispell*[1] (1993). These programs tolerate lower first guess accuracy by returning multiple guesses, allowing the user to make the final choice of the intended word. In contrast, some applications will require fully automatic correction for general-purpose texts (Kukich 1992).

It is clear that context-sensitive spelling correction offers better results than isolated-word error correction. The underlying task is to determine the relative degree of well formedness among alternative sentences (Mays et al. 1991). The question is what kind of knowledge (lexical, syntactic, semantic, ...) should be represented, utilised and combined to aid in this determination.

This study relies on the integrated use of three kinds of knowledge (syntagmatic, paradigmatic and statistical) in order to improve first guess accuracy in non-word context-sensitive correction for general unrestricted texts. Our techniques were applied to the corrections posed by *ispell*. Constraint Grammar (Karlsson et al. 1995) was chosen to represent syntagmatic knowledge. Its use as a part of speech tagger for English has been highly successful. Conceptual Density (Agirre and Rigau 1996) is the paradigmatic component chosen to discriminate semantically among potential noun corrections. This technique measures "affinity distance" between nouns using Wordnet (Miller 1990). Finally, general and document word-occurrence frequency-rates complete the set of knowledge sources combined. We knowingly did not use any model of common misspellings, the main reason being that we did not want to use knowledge about the error source. This work focuses on language models, not error models (typing errors, common misspellings, OCR mistakes, speech recognition mistakes, etc.).

The system was evaluated against two sets of texts: artificially generated errors from the Brown corpus (Francis and Kucera 1967) and genuine spelling errors from the Bank of English[2].

The remainder of this paper is organised as follows. Firstly, we present the techniques that

[1] *Ispell* was used for the spell-checking and correction candidate generation. Its assets include broad-coverage and excellent reliability.

[2] http://titania.cobuild.collins.co.uk/boe_info.html

will be evaluated and the way to combine them. Section 2 describes the experiments and shows the results, which are evaluated in section 3. Section 4 compares other relevant work in context sensitive correction.

1 The basic techniques

1.1 Constraint Grammar (CG)

Constraint Grammar was designed with the aim of being a language-independent and robust tool to disambiguate and analyse unrestricted texts. CG grammar statements are close to real text sentences and directly address parsing problems such as ambiguity. Its application to English (ENGCG[3]) resulted a very successful part of speech tagger for English. CG works on a text where all possible morphological interpretations have been assigned to each word-form by the ENGTWOL morphological analyser (Voutilainen and Heikkilä 1995). The role of CG is to apply a set of linguistic constraints that discard as many alternatives as possible, leaving at the end almost fully disambiguated sentences, with one morphological or syntactic interpretation for each word-form. The fact that CG tries to leave a unique interpretation for each word-form makes the formalism adequate to achieve our objective.

Application of Constraint Grammar

The text data was input to the morphological analyser. For each unrecognised word, *ispell* was applied, placing the morphological analyses of the correction proposals as alternative interpretations of the erroneous word (see example 1). EngCG-2 morphological disambiguation was applied to the resulting texts, ruling out the correction proposals with an incompatible POS (cf. example 2). We must note that the broad coverage lexicons of *ispell* and ENGTWOL are independent. This caused the correspondence between unknown words and *ispell*'s proposals not to be one to one with those of the EngCG-2 morphological analyser, especially in compound words. Such problems were solved considering that a word was correct if it was covered by any of the lexicons.

1.2 Conceptual Density (CD)

The discrimination of the correct category is unable to distinguish among readings belonging to the same category, so we also applied a word-sense disambiguator based on Wordnet, that had already been tried for nouns on free-running text. In our case it would choose the correction proposal semantically closer to the surrounding context. It has to be noticed that Conceptual Density can only be applied when all the proposals are categorised as nouns, due to the structure of Wordnet.

```
<our>
  "our" PRON PL ...
 <bos> ; INCORRECT OR SPELLING ERROR
  "boss" N S
  "boys" N P
  "bop" V S
  "Bose" <Proper>
```

Example 1. Proposals and morphological analysis for the misspelling *bos*

```
<our>
  "our" PRON PL ...
 <bos> ; INCORRECT OR SPELLING ERROR
  "boss" N S
  "boys" N P
  ~~"bop" V S~~
  "Bose" <Proper>
 <are>          ...
```

Example 2. CG leaves only nominal proposals

1.3 Frequency statistics (DF & BF)

Frequency data was calculated as word-form frequencies obtained from the document where the error was obtained (Document frequency, DF) or from the rest of the documents in the whole Brown Corpus (Brown frequency, BF). The experiments proved that word-forms were better suited for the task, compared to frequencies on lemmas.

1.4 Other interesting heuristics (H1, H2)

We eliminated proposals beginning with an uppercase character when the erroneous word did not begin with uppercase and there were alternative proposals beginning with lowercase. In example 1, the fourth reading for the misspelling "bos" was eliminated, as "Bose" would be at an editing distance of two from the misspelling (heuristic H1). This heuristic proved very reliable, and it was used in all experiments. After obtaining the first results, we also noticed that words with less than 4 characters like "si", "teh", ... (misspellings for "is" and "the") produced too many proposals, difficult to disambiguate. As they were one of the main error sources for our method, we also evaluated the results excluding them (heuristic H2).

[3] A recent version of ENGCG, known as EngCG-2, can be tested at http://www.conexor.fi/analysers.html

1.5 Combination of the basic techniques using votes

We considered all the possible combinations among the different techniques, e.g. CG+BF, BF+DF, and CG+DF. The weight of the vote can be varied for each technique, e.g. CG could have a weight of 2 and BF a weight of 1 (we will represent this combination as CG2+BF1). This would mean that the BF candidate(s) will only be chosen if CG does not select another option or if CG selects more than one proposal. Several combinations of weights were tried. This simple method to combine the techniques can be improved using optimization algorithms to choose the best weights among fractional values. Nevertheless, we did some trials weighting each technique with its expected precision, and no improvement was observed. As the best combination of techniques and weights for a given set of texts can vary, we separated the error corpora in two, trying all the possibilities on the first half, and testing the best ones on the second half (c.f. section 2.1).

2 The experiments

Based on each kind of knowledge, we built simple guessers and combined them in different ways. In the first phase, we evaluated all the possibilities and selected the best ones on part of the corpus with artificially generated errors. Finally, the best combinations were tested against the texts with genuine spelling errors.

2.1 The error corpora

We chose two different corpora for the experiment. The first one was obtained by systematically generating misspellings from a sample of the Brown Corpus, and the second one was a raw text with genuine errors. While the first one was ideal for experimenting, allowing for automatic verification, the second one offered a realistic setting. As we said before, we are testing language models, so that both kinds of data are appropriate. The corpora with artificial errors, artificial corpora for short, have the following features: a sample was extracted from SemCor (a subset of the Brown Corpus) selecting 150 paragraphs at random. This yielded a seed corpus of 505 sentences and 12659 tokens. To simulate spelling errors, a program named *antispell*, which applies Damerau's rules at random, was run, giving an average of one spelling error for each 20 words (non-words were left untouched). *Antispell* was run 8 times on the seed corpus, creating 8 different corpora with the same text but different errors. Nothing was done to prevent two errors in the same sentence, and some paragraphs did not have any error.

The corpus of genuine spelling errors, which we also call the "real" corpus for short, was magazine text from the Bank of English Corpus, which probably was not previously spell-checked (it contained many misspellings), so it was a good source of errors. Added to the difficulty of obtaining texts with real misspellings, there is the problem of marking the text and selecting the correct proposal for automatic evaluation.

As mentioned above, the artificial-error corpora were divided in two subsets. The first one was used for training purposes[4]. Both the second half and the "real" texts were used for testing.

2.2 Data for each corpora

The two corpora were passed trough *ispell*, and for each unknown word, all its correction proposals were inserted. Table 1 shows how, if the misspellings are generated at random, 23.5% of them are real words, and fall out of the scope of this work. Although we did not make a similar counting in the real texts, we observed that a similar percentage can be expected.

	1st half	2nd half	"real"
words	47584	47584	39733
errors	1772	1811	-[5]
non real-word errors	1354	1403	369
ispell proposals	7242	8083	1257
words with multiple proposals	810	852	158
long word errors (H2)	968	980	331
proposals for long words (H2)	2245	2313	807
long word errors (H2) with multiple proposals	430	425	124

Table 1. Number of errors and proposals

For the texts with genuine errors, the method used in the selection of the misspellings was the following: after applying *ispell*, no correction was found for 150 words (mainly proper nouns and foreign words), and there were about 300 which were formed by joining two consecutive words or by special affixation rules (*ispell* recognised them

[4] In fact, there is no training in the statistical sense. It just involves choosing the best alternatives for voting.

[5] As we focused on non-word words, there is not a count of real-word errors.

	Cover.%	Prec.%	#prop.
Basic techniques			
random baseline	100.00	54.36	1.00
random+H2	71.49	71.59	1.00
CG	99.85	86.91	2.33
CG+H2	71.42	95.86	1.70
BF	96.23	86.57	1.00
BF+H2	68.69	92.15	1.00
DF	90.55	89.97	1.02
DF+H2	62.92	96.13	1.01
CD	6.06	79.27	1.01
Combinations			
CG1+DF2	99.93	90.39	1.17
CG1+DF2+H2	71.49	96.38	1.12
CG1+DF1+BF1	99.93	89.14	1.03
CG1+DF1+BF1+H2	71.49	94.73	1.03
CG1+DF1+BF1+CD1	99.93	89.14	1.02
CG1+DF1+BF1+CD1+H2	71.49	94.63	1.02

Table 2. Results for several combinations (1st half)

	Cover.	Prec.	#prop
Basic techniques			
random baseline	100.00	23.70	1.00
random+H2	52.70	36.05	1.00
CG	99.75	78.09	3.23
CG+H2	52.57	90.68	2.58
BF	93.70	76.94	1.00
BF+H2	48.04	81.38	1.00
DF	84.20	81.96	1.03
DF+H2	38.48	89.49	1.03
CD	8.27	75.28	1.01
Combinations			
CG1+DF2	99.88	83.93	1.28
CG1+DF2+H2	52.70	91.86	1.43
CG1+DF1+BF1	99.88	81.83	1.04
CG1+DF1+BF1+H2	52.70	88.14	1.06
CG1+DF1+BF1+CD1	99.88	81.83	1.04
CG1+DF1+BF1+CD+H2	52.70	87.91	1.05

Table 3. Results on errors with multiple proposals (1st half)

correctly). This left 369 erroneous word-forms. After examining them we found that the correct word-form was among *ispell*'s proposals, with very few exceptions. Regarding the selection among the different alternatives for an erroneous word-form, we can see that around half of them has a single proposal. This gives a measure of the work to be done. For example, in the real error corpora, there were 158 word-forms with 1046 different proposals. This means an average of 6.62 proposals per word. If words of length less than 4 are not taken into account, there are 807 proposals, that is, 4.84 alternatives per word.

	Cover.%	Prec.%	#prop
Basic techniques			
random baseline	100.00	53.67	1.00
random+H2	69.85	71.53	1.00
DF	90.31	89.50	1.02
DF+H2	61.51	95.60	1.01
Combinations			
CG1+DF2	99.64	90.06	1.19
CG1+DF2+H2	69.85	95.71	1.22
CG1+DF1+BF1	99.64	87.77	1.03
CG1+DF1+BF1+H2	69.85	93.16	1.03
CG1+DF1+BF1+CD1	99.64	87.91	1.03
CG1+DF1+BF1+CD+H2	69.85	93.27	1.02

Table 4. Validation of the best combinations (2nd half)

	Cover.	Prec.	#pro
Basic techniques			
random baseline	100.00	23.71	1.00
random+H2	50.12	34.35	1.00
DF	84.04	81.42	1.03
DF+H2	36.32	87.66	1.04
Combinations			
CG1+DF2	99.41	83.59	1.31
CG1+DF2+H2	50.12	90.12	1.50
CG1+DF1+BF1	99.41	79.81	1.05
CG1+DF1+BF1+H2	50.12	84.24	1.06
CG1+DF1+BF1+CD1	99.41	80.05	1.05
CG1+DF1+BF1+CD1+H2	50.12	84.47	1.06

Table 5. Results on errors with multiple proposals (2nd half)

2.3 Results

We mainly considered three measures:

- coverage: the number of errors for which the technique yields an answer.
- precision: the number of errors with the correct proposal among the selected ones
- remaining proposals: the average number of selected proposals.

2.3.1 Search for the best combinations

Table 2 shows the results on the training corpora. We omit many combinations that we tried, for the sake of brevity. As a baseline, we show the results when the selection is done at random. Heuristic H1 is applied in all the cases, while tests are performed with and without heuristic H2. If we focus on the errors for which *ispell* generates more than one correction proposal (cf. table 3), we get a better estimate of the contribution of each guesser. There were 8.26 proposals per word in the general

	Cover. %	Prec. %	#prop.
Basic techniques			
random baseline	100.00	69.92	1.00
random+H2	89.70	75.47	1.00
CG	99.19	84.15	1.61
CG+H2	89.43	90.30	1.57
DF	70.19	93.05	1.02
DF+H2	61.52	97.80	1.00
BF	98.37	80.99	1.00
BF+H2	88.08	85.54	1.00
Combinations			
CG1+DF2	100.00	87.26	1.42
CG1+DF2+H2	89.70	90.94	1.43
CG1+DF1+BF1	100.00	80.76	1.02
CG1+DF1+BF1+H2	89.70	84.89	1.02

Table 6. Best combinations ("real" corpus)

	Cover. %	Prec. %	#prop
Basic techniques			
random baseline	100.00	29.75	1.00
random+H2	76.54	34.52	1.00
CG	98.10	62.58	2.45
CG+H2	75.93	73.98	2.52
DF	30.38	62.50	1.13
DF+H2	12.35	75.00	1.05
BF	96.20	54.61	1.00
BF+H2	72.84	60.17	1.00
Combinations			
CG1+DF2	100.00	70.25	1.99
CG1+DF2+H2	76.24	75.81	2.15
CG1+DF1+BF1	100.00	55.06	1.04
CG1+DF1+BF1+H2	76.54	59.68	1.05

Table 7. Results on errors with multiple proposals ("real" corpus)

case, and 3.96 when H2 is applied. The results for all the techniques are well above the random baseline. The single best techniques are DF and CG. CG shows good results on precision, but fails to choose a single proposal. H2 raises the precision of all techniques at the cost of losing coverage. CD is the weakest of all techniques, and we did not test it with the other corpora. Regarding the combinations, CG1+DF2+H2 gets the best precision overall, but it only gets 52% coverage, with 1.43 remaining proposals. Nearly 100% coverage is attained by the H2 combinations, with highest precision for CG1+DF2 (83% precision, 1.28 proposals).

2.3.2 Validation of the best combinations

In the second phase, we evaluated the best combinations on another corpus with artificial errors. Tables 4 and 5 show the results, which agree with those obtained in 2.3.1. They show slightly lower percentages but always in parallel.

2.3.3 Corpus of genuine errors

As a final step we evaluated the best combinations on the corpus with genuine typing errors. Table 6 shows the overall results obtained, and table 7 the results for errors with multiple proposals. For the latter there were 6.62 proposals per word in the general case (2 less than in the artificial corpus), and 4.84 when heuristic H2 is applied (one more that in the artificial corpus). These tables are further commented in the following section.

3 Evaluation of results

This section reviews the results obtained. The results for the "real" corpus are evaluated first, and the comparison with the other corpora comes later. Concerning the application of each of the simple techniques separately[6]:

- Any of the guessers performs much better than random.
- DF has a high precision (75%) at the cost of a low coverage (12%). The difference in coverage compared to the artificial error corpora (84%) is mainly due to the smaller size of the documents in the real error corpus (around 50 words per document). For medium-sized documents we expect a coverage similar to that of the artificial error corpora.
- BF offers lower precision (54%) with the gains of a broad coverage (96%).
- CG presents 62% precision with nearly 100% coverage, but at the cost of leaving many proposals (2.45)
- The use of CD works only with a small fraction of the errors giving modest results. The fact that it was only applied a few times prevents us from making further conclusions.

Combining the techniques, the results improve:

- The CG1+DF2 combination offers the best results in coverage (100%) and precision (70%) for all tests. As can be seen, CG raises the coverage of the DF method, at the cost of also increasing the number of proposals (1.9) per erroneous word. Had the coverage of DF increased, so would also the number of

[6] If not explicitly noted, the figures and comments refer to the "real" corpus, table 7.

proposals decrease for this combination, for instance, close to that of the artificial error corpora (1.28).

- The CG1+DF1+BF1 combination provides the same coverage with nearly one interpretation per word, but decreasing precision to a 55%.
- If full coverage is not necessary, the use of the H2 heuristic raises the precision at least 4% for all combinations.

When comparing these results with those of the artificial errors, the precisions in tables 2, 4 and 6 can be misleading. The reason is that the coverage of some techniques varies and the precision varies accordingly. For instance, coverage of DF is around 70% for real errors and 90% for artificial errors, while precisions are 93% and 89% respectively (cf. tables 6 and 2). This increase in precision is not due to the better performance of DF[7], but can be explained because the lower the coverage, the higher the proportion of errors with a single proposal, and therefore the higher the precision.

The comparison between tables 3 and 7 is more clarifying. The performance of all techniques drops in table 7. Precision of CG and BF drops 15 and 20 points. DF goes down 20 points in precision and 50 points in coverage. This latter degradation is not surprising, as the length of the documents in this corpus is only of 50 words on average. Had we had access to medium sized documents, we would expect a coverage similar to that of the artificial error corpora.

The best combinations hold for the "real" texts, as before. The highest precision is for CG1+DF2 (with and without H2). The number of proposals left is higher in the "real" texts than in the artificial ones (1.99 to 1.28). It can be explained because DF does not manage to cover all errors, and that leaves many CG proposals untouched.

We think that the drop in performance for the "real" texts was caused by different factors. First of all, we already mentioned that the size of the documents strongly affected DF. Secondly, the nature of the errors changes: the algorithm to produce spelling errors was biased in favour of frequent words, mostly short ones. We will have to analyse this question further, specially regarding the origin of the natural errors. Lastly, BF was trained on the Brown corpus on American English, while the "real" texts come from the Bank of English. Presumably, this could have also affected negatively the performance of these algorithms.

Back to table 6, the figures reveal which would be the output of the correction system. Either we get a single proposal 98% of the times (1.02 proposals left on average) with 80% precision for all non-word errors in the text (CG1+DF1+BF1) or we can get a higher precision of 90% with 89% coverage and an average of 1.43 proposals (CG1+DF2+H2).

4 Comparison with other context-sensitive correction systems

There is not much literature about automatic spelling correction with a single proposal. Menezo et al. (1996) present a spelling/grammar checker that adjusts its strategy dynamically taking into account different lexical agents (dictionaries, ...), the user and the kind of text. Although no quantitative results are given, this is in accord with using document and general frequencies.

Mays et al. (1991) present the initial success of applying word trigram conditional probabilities to the problem of context based detection and correction of real-word errors.

Yarowsky (1994) experiments with the use of decision lists for lexical ambiguity resolution, using context features like local syntactic patterns and collocational information, so that multiple types of evidence are considered in the context of an ambiguous word. In addition to word-forms, the patterns involve POS tags and lemmas. The algorithm is evaluated in missing accent restoration task for Spanish and French text, against a predefined set of a few words giving an accuracy over 99%.

Golding and Schabes (1996) propose a hybrid method that combines part-of-speech trigrams and context features in order to detect and correct real-word errors. They present an experiment where their system has substantially higher performance than the grammar checker in MS Word, but its coverage is limited to eighteen particular confusion sets composed by two or three similar words (e.g.: weather, whether).

The last three systems rely on a previously collected set of confusion sets (sets of similar words or accentuation ambiguities). On the contrary, our system has to choose a single

[7] In fact the contrary is deduced from tables 3 and 7.

proposal for any possible spelling error, and it is therefore impossible to collect the confusion sets (i.e. sets of proposals for each spelling error) beforehand. We also need to correct as many errors as possible, even if the amount of data for a particular case is scarce.

Conclusion

This work presents a study of different methods that build on the correction proposals of *ispell*, aiming at giving a single correction proposal for misspellings. One of the difficult aspects of the problem is that of testing the results. For that reason, we used both a corpus with artificially generated errors for training and testing, and a corpus with genuine errors for testing.

Examining the results, we observe that the results improve as more context is taken into account. The word-form frequencies serve as a crude but helpful criterion for choosing the correct proposal. The precision increases as closer contexts, like document frequencies and Constraint Grammar are incorporated. From the results on the corpus of genuine errors we can conclude the following. Firstly, the correct word is among *ispell*'s proposals 100% of the times, which means that all errors can be recovered. Secondly, the expected output from our present system is that it will correct automatically the spelling errors with either 80% precision with full coverage or 90% precision with 89% coverage and leaving an average of 1.43 proposals.

Two of the techniques proposed, Brown Frequencies and Conceptual Density, did not yield useful results. CD only works for a very small fraction of the errors, which prevents us from making further conclusions.

There are reasons to expect better results in the future. First of all, the corpus with genuine errors contained very short documents, which caused the performance of DF to degrade substantially. Further tests with longer documents should yield better results. Secondly, we collected frequencies from an American English corpus to correct British English texts. Once this language mismatch is solved, better performance should be obtained. Lastly, there is room for improvement in the techniques themselves. We knowingly did not use any model of common misspellings. Although we expect limited improvement, stronger methods to combine the techniques can also be tried.

Continuing with our goal of attaining a single proposal as reliably as possible, we will focus on short words and we plan to also include more syntactic and semantic context in the process by means of collocational information. This step opens different questions about the size of the corpora needed for accessing the data and the space needed to store the information.

Acknowledgements

This research was supported by the Basque Government, the University of the Basque Country and the CICYT (Comisión Interministerial de Ciencia y Tecnología).

References

Agirre E. and Rigau G. (1996) *Word sense disambiguation using conceptual density.* In Proc. of COLING-96, Copenhagen, Denmark.

Golding A. and Schabes. Y. (1996) *Combining trigram-based and feature-based methods for context-sensitive spelling correction.* In Proc. of the 34th ACL Meeting, Santa Cruz, CA.

Ispell (1993) International Ispell Version 3.1.00, 10/08/93.

Francis S.and Kucera H. (1967) *Computing Analysis of Present-Day American English*. Brown Univ. Press.

Karlsson F., Voutilainen A., Heikkilä J. and Anttila A. (1995) *Constraint Grammar: a Language Independent System for Parsing Unrestricted Text.* Ed.Mouton de Gruyter.

Koskenniemi K. (1983) *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production.* University of Helsinki.

Kukich K. (1992) *Techniques for automatically correcting words in text.* In ACM Computing Surveys, Vol. 24, N. 4, December, pp. 377-439.

Mays E., Damerau F. and Mercer. R. (1991) *Context based spelling correction* . Information Processing & Management, Vol. 27, N. 5, pp. 517-522.

Miller G. (1990) *Five papers on WordNet.* Special Issue of the Int. Journal of Lexicography, Vol. 3, N. 4.

Menezo J., Genthial D. and Courtin J. (1996) Reconnaisances pluri-lexicales dans CELINE, un système multi-agents de détection et correction des erreurs. NLP + IA 96, Moncton, N. B., Canada.

Yarowsky D. (1994) *Decision lists for lexical ambiguity resolution.* In Proceedings of the 32nd ACL Meeting, Las Cruces, NM, pp.88-95.

"Towards a single proosal in spelling sorrection"

Eneko Agirre, Koldo Gojenola
Kepa Sarasola, Atro Voutilainen

UPV-EHU / LSI / TR 8-98

Title:

Towards a single proposal in spelling correction

Authors:

Eneko Agirre, Koldo Gojenola, Kepa Sarasola
Dept. of Computer Languages and Systems
University of the Basque Country, 649 P. K.,
E-20080 Donostia, Basque Country
eneko@si.ehu.es

Atro Voutilainen
Department of General Linguistics
University of Helsinki, P.O. Box 4
FIN-00014 Helsinki, Finland
avoutila@ling.helsinki.fi

Abstract:

The study here presented relies on the integrated use of three kinds of knowledge (syntagmatic, paradigmatic and statistical) in order to improve first-guess accuracy in non-word context-sensitive correction for general unrestricted texts. State of the art spelling correction systems, e.g. *ispell*, in addition to detecting spelling errors also assist the user by offering a set of candidate corrections that are close to the misspelled word. Based on the correction proposals of *ispell*, we built several guessers which were combined in different ways. Firstly, we evaluated all the possibilities and selected the best ones on a corpus with artificially generated typing errors. Secondly, the best combinations were tested on texts containing genuine spelling errors. The results for the latter suggest that we can expect automatic non-word correction for *all* the errors in a free-running text with 90% precision and a single proposal 24 times out of 25 (1.04 proposals on average).

Topic areas:

spelling correction

Towards a single proposal in spelling correction

Abstract

The study here presented relies on the integrated use of three kinds of knowledge (syntagmatic, paradigmatic and statistical) in order to improve first-guess accuracy in non-word context-sensitive correction for general unrestricted texts. State of the art spelling correction systems, e.g. *ispell*, in addition to detecting spelling errors also assist the user by offering a set of candidate corrections that are close to the misspelled word. Based on the correction proposals of *ispell*, we built several guessers which were combined in different ways. Firstly, we evaluated all the possibilities and selected the best ones on a corpus with artificially generated typing errors. Secondly, the best combinations were tested on texts containing genuine spelling errors. The results for the latter suggest that we can expect automatic non-word correction for *all* the errors in a free-running text with 90% precision and a single proposal 24 times out of 25 (1.04 proposals on average).

Introduction

The problem of devising algorithms and techniques for automatically correcting words in text remains being a research challenge. Existing spelling correction techniques are limited in their scope and accuracy. In addition to detecting spelling errors many programs assist users by offering a set of candidate corrections that are close to the misspelled word. This is true for most of the commercial word-processors as well as the Unix-based spelling-corrector *ispell*[1] (1993). These programs tolerate lower first guess accuracy by returning multiple guesses and allowing the user to make the final choice of the intended word. In contrast, some applications will require fully automatic correction for general purpose texts (Kukich 1992).

It is clear that context-sensitive correction will offer better results than isolated-word error correction. The task underlying context-sensitive spelling correction is to determine the relative degree of well-formedness among alternative sentences (Mays et al. 1991). The question is what kind of knowledge (lexical, syntactic, semantic, statistical, ...) should be represented, utilised and combined to aid in this determination.

The study here presented relies on the integrated use of three kinds of knowledge (syntagmatic, paradigmatic and statistical) in order to improve first guess accuracy in nonword context-sensitive correction for general unrestricted texts. Our techniques were applied on the corrections posed by *ispell*. Constraint Grammar (Karlsson 1995) was chosen to represent syntagmatic knowledge. Its use as a part of speech tagger for English was completely successful. Conceptual Density (Agirre and Rigau 1997) is the paradigmatic component chosen to discriminate semantically among potential noun corrections. This technique measures "affinity distance" between nouns using Wordnet (Miller 1990). Information on affinity to context was also collected from corpora, in the form of collocational and cooccurrence statistical features (Yarowsky 1994). Finally, general and document word-occurrence frequency-rates complete the set of different knowledge sources combined in the system. We knowingly did not use any model of common misspellings, the main reason being that we did not want to use knowledge about the error source. This work focuses on language models, not error models (typing errors, common misspellings, OCR mistakes, speech recognition mistakes, etc.).

The system was evaluated on two sets of texts: artificially generated typing errors from the

[1] Ispell was used for the spell-checking and correction candidate generation. Its assets include broad-coverage, excellent reliability (cf. the conclusion) and the fact that it is able to produce several kinds of output, e.g. errors and proposals only.

Brown corpus (Francis & Kucera 1967) and genuine spelling errors from the Bank of English[2].

The remainder of this paper is organised as follows. Firstly, we present the techniques that will be evaluated and the way to combine them. Section 2 describes the experiments performed and shows the results, which are evaluated in section 3. Section 4 compares other relevant work in context sensitive correction. Finally, the paper ends with some concluding remarks.

1 The basic techniques

1.1 Constraint Grammar (CG)

Constraint Grammar was designed with the aim of being a language-independent and robust tool to disambiguate and analyse unrestricted texts. The CG grammar statements are close to real text sentences and directly address some notorious parsing problems, especially ambiguity. Its application to English (ENGCG) is a very successful part of speech tagger for English.

These are four major steps in the CG morphosyntactic treatment of texts: morphological analysis, morphological disambiguation, determination of clause boundaries and the assignment of syntactic functions. CG works on a text where all the possible morphological interpretations have been assigned to each word-form by the ENGTWOL morphological analyser (Koskenniemi 1983). The basic parsing strategy is to profit from the existing morphological information. Every relevant structure is assigned directly via lexicon, morphology and mappings from morphology to syntax. The role of CG is to apply a set of linguistic constraints that discard as many alternatives as possible, leaving at the end almost fully disambiguated sentences, with one morphological/syntactic interpretation for each word-form. The fact that CG tries to leave a unique morphological/syntactic interpretation for each word-form makes this formalism adequate to achieve our objective.

Application of Constraint Grammar

The text data was input to the morphological analyser (ENGTWOL). For each unrecognised word *ispell* was applied, placing the morphological analyses of the correction proposals as alternative interpretations of the erroneous word (see Example 1).

```
<our>
  "our" PRON PL ...
<bos> ; INCORRECT OR SPELLING ERROR
  "boss" N S
  "boys" N P
  "bop"  V S
  "Bose" <Proper>
<are>          ...
```

Example 1.
Proposals and morphological analysis for the misspelling *bos*.

The CG morphological disambiguation was applied on the resulting texts, ruling out the correction proposals with an incompatible POS (cf. example 2).

```
<our>
  "our" PRON PL ...
<bos> ; INCORRECT OR SPELLING ERROR
  "boss" N S
  "boys" N P
  ~~"bop"  V S~~
  "Bose" <Proper>
<are>          ...
```

Example 2.
CG leaves only nominal proposals.

We have to note that the broad coverage lexicons of *ispell* and ENGTWOL are independent. This caused the correspondence between unknown words and the proposals given by *ispell* not to be one to one with those of the ENGTWOL lexicon, especially in compound words. Such problems were solved considering that a word was correct if it was covered by any of the lexicons.

1.2 Conceptual Density (CD)

The discrimination of the correct category is unable to distinguish among readings belonging to the same category, so that we also applied a

[2] http://titania.cobuild.collins.co.uk/boe_info.html

word-sense disambiguator (Agirre & Rigau 1996) based on Wordnet to this task. The word-sense disambiguator had already been tried for nouns on free-running text. In our case the disambiguator would choose the correction proposal semantically closer to the surrounding context. It has to be noted that Conceptual Density can only be applied whenever all the proposals are categorised as nouns.

1.3 Frequency statistics (DF & BF)

Frequency data was calculated as word-form frequencies obtained from the document where the error was obtained (Document frequency, DF) or from the rest of the documents in the whole Brown Corpus (Brown frequency, BF). The experiments proved that word-forms were better suited for the task, compared to frequencies on lemmas.

1.4 Context statistics (CX)

In accordance to the proposals of (Yarowsky 1994), we modelled the lexical preference of the proposals. Context features were collected for all the words in the Brown Corpus (minus the test documents). The collected features were:

- word bigrams
- word trigrams
- context words in a ±20 word-window

When processing an error, the features for each proposal were retrieved and their weight measured using log-likelihood (Yarowsky 1994). The proposal with the strongest feature would be chosen, under the supposition that it would be the best fitted for the context of the error.

1.5 Other interesting heuristics (H1, H2)

We eliminated proposals beginning with an uppercase character when the erroneous word did not begin with an uppercase letter and there were alternative proposals beginning with lowercase. In example 1 of the previous section, the fourth reading for the misspelling "bos" was eliminated, as "Bose" would be at an editing distance of two from the misspelling (heuristic H1). This heuristic proved very reliable, and it was used in all experiments.

After obtaining the first results, we also noted that words with less than 4 characters like "si", "teh", ... (misspellings for "is" and "the") produced too many proposals, difficult to disambiguate. As they were one of the main error sources for our method, we also evaluated the results excluding them (heuristic H2).

1.6 Combination of the basic techniques using votes

We considered all the possible combinations among the different techniques e.g. CG+BF, BF+DF, CG+DF+CX, etc.

The weight of the vote can be varied for each technique, e.g. CG could have a weight of 2 and BF a weight of 1 (we will represent this combination as CG2+BF1). This would mean that the BF candidate(s) will only be chosen if CG does not select another option. Several combinations of weights were tried.

As the best combination of techniques and weights for a given set of texts can vary we separated the error corpora in two, trying all the possibilities on the first half, and testing the performance of the best ones on the second half (c.f. section 2.1).

This simple method to combine the techniques can be improved using optimization algorithms to choose the best weigths among fractional values. Nevertheless, we did some trials weighting each technique with its expected precision and no imprevement was observed.

2 The experiments

Based on each kind of knowledge we built a simple guesser, and combined them in different ways. In a first phase, we evaluated all the possibilities and selected the best ones on a part of the corpus with artificially generated typing errors. Finally, the best combinations were tested on the texts with genuine spelling errors.

2.1 The error corpora

As we have explained before, we chose two different corpora for the experiment. The first one was obtained by systematically generating misspellings from a sample of the Brown Corpus, and the second one was a raw text with genuine errors. While the first one was ideal for experimenting with different parameters, allowing for automatic verification, the second offered a realistic setting.

The corpora with artificial errors, artificial corpora for short, have the following features: a sample was extracted from SemCor (a subset of the Brown Corpus) selecting 150 paragraphs at random. This yielded a seed corpus of 505 sentences and 12659 tokens. To simulate spelling errors a program named *antispell* which applies Damerau's rules at random was run, creating an average of one spelling error for each 20 words (nonwords were left untouched). *Antispell* was run 8 times on the seed corpus, creating 8 different corpus with the same text but different errors. Nothing was done to prevent two errors in the same sentence, and some paragraphs did not have any error.

The corpus of genuine spelling errors, which we also call the "real" corpus for short, was magazine text from the Bank of English Corpus, which was not previously spell-checked. Added to the difficulty of obtaining texts with real misspellings there is the problem of marking the text and selecting the correct proposal for automatic evaluation.

As mentioned above, the artificial-error corpora were divided in two subsets. The first one is composed of the first half, i.e. sets 1, 2, 3 and 4. It was used for training purposes[3]. The second half comprises texts 5, 6, 7 and 8. Both the second half and the "real" texts were used for testing.

2.2 Data for each corpora

The two corpora were passed trough *ispell*, and for each unknown word all its correction proposals were inserted.

Table 1 shows how, if the misspellings are generated at random, 23.5% of them are real words, and fall out of the scope of this work. Although we did not made a similar counting in the real texts, we observed that a similar percentage can be expected.

	1st half	2nd half	"real"
words	47584	47584	39733
errors	1772	1811	-[4]
non real-word errors	1354	1403	369
ispell proposals	7242	8083	1257
words with multiple proposals	810	852	158
long word errors (H2)	968	980	331
proposals for long words (H2)	2245	2313	807
long word errors (H2) with multiple proposals	430	425	124

Table 1. Number of errors and proposals

For the texts with genuine errors, the method used in the selection of the misspellings was the following: after applying *ispell*, no correction was found for 150 words (mainly proper nouns and foreign words), and there were about 300 which were formed by joining two consecutive words or by special affixation rules (*ispell* recognised them correctly most of the times). This left 369 erroneous word-forms. After examining them we found that the correct word-form was, with very few exceptions, among *ispell*'s proposals.

Regarding the selection among the different alternatives for an erroneous word-form, we see that around half of them have a single proposal. This gives a measure of the work to be done. For example, in the real error corpora, there were 158 word-forms with 1046 different proposals. This means an average of 6.62 proposals per word. If words of length less than 4 are not taken into account, there are 807 proposals, that is, 4.84 alternatives per word.

[3] In fact, there is no training in the statistical sense, but it is rather choosing the best alternatives for voting (cf. 1.6).

[4] As we focused on unknown words, there is not a count of real-word errors.

	cover. %	prec. %	#prop.
Basic techniques			
random baseline	100.00	54.36	1.00
random+H2	71.49	71.59	1.00
CG	99.85	86.91	2.33
CG+H2	71.42	95.86	1.70
BF	96.23	86.57	1.00
BF+H2	68.69	92.15	1.00
DF	90.55	89.97	1.02
DF+H2	62.92	96.13	1.01
CX	96.70	91.20	1.01
CX+H2	68.54	95.70	1.01
CD	6.06	79.27	1.01
Combinations			
CG1+DF2	99.93	90.39	1.17
CG1+DF2+H2	71.49	96.38	1.12
CG1+DF1+BF1	99.93	89.14	1.03
CG1+DF1+BF1+H2	71.49	94.73	1.03
CG1+DF1+BF1+CD1	99.93	89.14	1.02
CG1+DF1+BF1+CD1+H2	71.49	94.63	1.02
CG1+DF1+CX1	99.93	91.90	1.07
CG1+DF1+CX1+H2	71.49	96.50	1.05
CG1+DF1+CX2	99.93	91.30	1.04
CG1+DF1+CX2+H2	71.49	95.70	1.04

Table 2. Results for several combinations (1st half)

	Cover. %	Prec. %	#prop
Basic techniques			
random baseline	100.00	23.70	1.00
random+H2	52.70	36.05	1.00
CG	99.75	78.09	3.23
CG+H2	52.57	90.68	2.58
BF	93.70	76.94	1.00
BF+H2	48.04	81.38	1.00
DF	84.20	81.96	1.03
DF H2	38.48	89.49	1.03
CX	94.48	84.94	1.02
CX+H2	47.79	89.77	1.02
CD	8.27	75.28	1.01
Combinations			
CG1+DF2	99.88	83.93	1.28
CG1+DF2+H2	52.70	91.86	1.43
CG1+DF1+BF1	99.88	81.83	1.04
CG1+DF1+BF1+H2	52.70	88.14	1.06
CG1+DF1+BF1+CD1	99.88	81.83	1.04
CG1+DF1+BF1+CD+H2	52.70	87.91	1.05
CG1+DF1+CX1	99.88	86.45	1.12
CG1+DF1+CX1+H2	52.70	92.12	1.11
CG1+DF1+CX2	99.88	85.45	1.07
CG1+DF1+CX2+H2	52.70	90.32	1.09

Table 3. Results on errors with multiple proposals (1st half)

	cover. %	Prec.%	#prop.
Basic techniques			
Random baseline	100.00	53.67	1.00
Random+H2	69.85	71.53	1.00
DF	90.31	89.50	1.02
DF H2	61.51	95.60	1.01
CX	97.20	91.00	1.01
CX+H2	67.93	94.30	1.01
Combinations			
CG1+DF2	99.64	90.06	1.19
CG1+DF2+H2	69.85	95.71	1.22
CG1+DF1+BF1	99.64	87.77	1.03
CG1+DF1+BF1+H2	69.85	93.16	1.03
CG1+DF1+BF1+CD1	99.64	87.91	1.03
CG1+DF1+BF1+CD+H2	69.85	93.27	1.02
CG1+DF1+CX1	99.71	91.60	1.09
CG1+DF1+CX1+H2	69.85	95.10	1.07
CG1+DF1+CX2	99.71	91.07	1.04
CG1+DF1+CX2+H2	69.85	94.18	1.03

Table 4. Validation of the best combinations (2nd half)

	Cover. %	Prec. %	#prop
Basic techniques			
random baseline	100.00	23.71	1.00
random+H2	50.12	34.35	1.00
DF	84.04	81.42	1.03
DF H2	36.32	87.66	1.04
CX	95.39	84.90	1.02
CX H2	46.93	86.35	1.02
Combinations			
CG1+DF2	99.41	83.59	1.31
CG1+DF2+H2	50.12	90.12	1.50
CG1+DF1+BF1	99.41	79.81	1.05
CG1+DF1+BF1+H2	50.12	84.24	1.06
CG1+DF1+BF1+CD1	99.41	80.05	1.05
CG1+DF1+BF1+CD+H2	50.12	84.47	1.06
CG1+DF1+CX1	99.53	86.14	1.15
CG1+DF1+CX1+H2	50.12	88.70	1.16
CG1+DF1+CX2	99.53	85.26	1.07
CG1+DF1+CX2+H2	50.12	86.59	1.07

Table 5. Results on errors with multiple proposals (2nd half)

2.3 Results

There are three measures which we deemed important:

- coverage: the number of errors for which the technique yields an answer.
- precision: the number of errors for which the correct proposal remains among the selected ones

- remaining proposals: the average number of selected proposals.

2.3.1 Search for the best combinations

Table 2 shows some of the results obtained for the training corpora (1st half of the corpora with artificial errors), with the most interesting results shadowed. We omit most of the combinations we tried for the sake of brevity. As a baseline, we show the results when the selection is done at random. Heuristic H1 is applied in all of the cases, while tests are performed with and without heuristic H2.

If we focus on the errors for which *ispell* generates more than one correction proposal (cf. table 3), we can get a better estimate of the contribution of each guesser. There were 8.26 proposals per word in the general case, and 3.96 when heuristic H2 is applied. The results for all the techniques are well above the random baseline. The single best techniques are DF and CX. CG has also good results on precision, but fails to choose a single proposal. The H2 heuristic raises the precision of all techniques at least 5 points, at the cost of losing coverage. CD is the weakest of all techniques, and we did not test it with the other corpora.

Regarding the combinations, CG+DF+CX+H2 gets the best precision overall, but only gets 52% coverage. CG1+DF2+H2 follows close, with more proposals. Nearly 100% coverage is attained by the combinations without H2, with highest precision for CG+DF+CX (86% precision, 1.12 proposals). If CX gets double votes (CG1+DF1+CX2) fewer proposals are selected (1.07) but one point is lost in precision.

2.3.2 Validation of the best combinations

In the second phase, we evaluated the best combinations on another corpus with artificially generated typing errors. Tables 4 and 5 show that the results for the 2nd half agree with those obtained in 2.3.1. The results show slightly lower percentages for all techniques, but always in parallel. This confirms that the best combinations hold for other texts.

	Cover. %	prec. %	#prop.
Basic techniques			
random baseline	100.00	69.92	1.00
random+H2	89.70	75.47	1.00
CG	99.19	84.15	1.61
CG+H2	89.43	90.30	1.57
DF	70.19	93.05	1.02
DF+H2	61.52	97.80	1.00
BF	98.37	80.99	1.00
BF+H2	88.08	85.54	1.00
CX	97.02	89.10	1.02
CX+H2	85.64	91.50	1.01
Combinations			
CG1+DF2	100.00	87.26	1.42
CG1+DF2+H2	89.70	90.94	1.43
CG1+DF1+BF1	100.00	80.76	1.02
CG1+DF1+BF1+H2	89.70	84.89	1.02
CG1+DF1+CX1	100.00	90.80	1.24
CG1+DF1+CX1+H2	89.70	93.10	1.20
CG1+DF1+CX2	100.00	89.70	1.04
CG1+DF1+CX2+H2	89.70	91.80	1.03

Table 6. Best combinations ("real" corpus)

	cover. %	prec. %	#prop
Basic techniques			
random baseline	100.00	29.75	1.00
random+H2	76.54	34.52	1.00
CG	98.10	62.58	2.45
CG+H2	75.93	73.98	2.52
DF	30.38	62.50	1.13
DF+H2	12.35	75.00	1.05
BF	96.20	54.61	1.00
BF+H2	72.84	60.17	1.00
CX	93.21	74.16	1.05
CX+H2	67.28	75.36	1.03
Combinations			
CG1+DF2	100.00	70.25	1.99
CG1+DF2+H2	76.24	75.81	2.15
CG1+DF1+BF1	100.00	55.06	1.04
CG1+DF1+BF1+H2	76.54	59.68	1.05
CG1+DF1+CX1	100.00	78.51	1.56
CG1+DF1+CX1+H2	76.54	81.58	1.53
CG1+DF1+CX2	100.00	75.94	1.09
CG1+DF1+CX2+H2	76.54	78.11	1.08

Table 7. Results on errors with multiple proposals ("real" corpus)

2.3.3 Corpus of genuine errors

As a final step we evaluated the best combinations on the corpus with genuine typing errors. Table 6 shows the overall results obtained, and table 7 the results for errors with multiple proposals. For the latter there were 6.62

proposals per word in the general case (2 less than in the artificial corpus), and 4.84 when heuristic H2 is applied (one more that in the artificial corpus).

These tables are further commented in the following section.

3 Evaluation of results

This sections reviews the results obtained. The results for the "real" texts are evaluated first, and the comparison with the other texts comes later.

Concerning the application of each of the simple techniques separately[5]:

- Any of the guessers performs much better than random.
- CX has the highest precision (74%) with 93% coverage.
- DF has lower precision (62%) and lower coverage (30%).
- BF offers lower precision (54%) with the gains of a broad coverage (96%).
- CG presents 62% precision with nearly 100% coverage, but at the cost of leaving many proposals (2.45)

When the techniques are combined, the results improve:

- The CG+DF+CX combination offers the best results in coverage (close to 100%) and precision for all tests (78% in table 7).
- If CX gets double weight, CG1+DF1+CX2, some precision is lost (76%), but the number of proposals left is more satisfactory (1.09 against 1.56).
- CG1+DF2 attains 70% precision. As it can be seen, CG raises the coverage of the DF method, at the cost of also increasing the number of proposals (1.9) per erroneous word. Had the coverage of DF increased, so would also decrease the number of proposals for this combination, for instance, close to that of the artificial error corpora (1.28).
- The CG1+DF1+BF1 combination provides the same coverage with nearly one interpretation per word, but decreasing precision to a 55%.
- If full coverage is not necessary, the use of the H2 heuristic raises the precision at least 3% for all combinations.

When comparing these results with those of the artificial errors, the precisions in tables 2, 4 and 6 can be misleading. The reason is that the coverage of some techniques varies and the precision varies accordingly. For instance, coverage of DF is around 70% for real errors and 90% for artificial errors, while precisions are 93% and 89% respectively (cf. tables 6 and 2). This raise in precision is not due to the better performance of DF[6], but can be explained because the lower the coverage the higher the proportion of errors with a single proposal, and therefore the higher the precision.

The comparison between tables 3 and 7 is more clarifying. The performance of all techniques drops in table 7. Precision of CG, CX and BF drops 15, 10 and 20 points respectively. DF goes down 20 points in precision and 50 points in coverage. This latter degradation in performance is not surprising, as the length of the documents in this corpus is only of 50 words on average. Had we used medium sized documents, we would expect a coverage similar to that of the artificial error corpora.

The best combinations hold for the "real" texts, as before. The highest precision is for CG+DF+CX (with and without H2). The number of proposals left is higher in the "real" texts than in the artificial texts (1.56 to 1.12). This can be explained because DF and CX do not manage to cover all errors, and that leaves many proposals of CG untouched.

We think that the drop in performance for the "real" texts was caused by different factors. First of all, we already mentioned that the size of the documents strongly affected DF. Secondly, the nature of the errors change: the algorithm to

[5] If not explicitly noted, the figures and comments refer to the "real" text, table 7.

[6] In fact the contrary is deduced from the data in tables 3 and 7.

produce spelling errors was biased in favour of frequent words, mostly short ones. We will have to analyze this question further, specially regarding the origin of the natural errors. Lastly, two techniques, namely BF and CX, were trained on the Brown corpus on American English, while the "real" texts come from the Bank of English. Presumably, this could have also affected negatively the performance of these algorithms.

Back to table 6, the figures reveal which would be the output of the correction system. Either we get a single proposal 24 times out of 25 (1.04 proposals left on average) with 90% precision for all non-word errors in the text (CG1+DF1+CX2) or we can get a higher precision of 93% with 90% coverage and an average of 1.20 proposals (CG+DF+CX+H2).

4 Comparing with other context-sensitive correction systems

There is not much literature about automatic spelling correction with a single proposal. Menezo et al. (1996) present the design of an interactive automatic spelling and grammar checker/corrector based on an architecture of distributed artificial intelligence and a multi-agent system. It allows to adjust its strategy dynamically taking into account the different lexical agents (dictionaries, ...), the user, the kind of text, and even the window. Although no quantitative results are given, this is in accord with using the document and general frequencies.

Mays et al. (1991) present the initial success of applying word trigram conditional probabilities to the problem of context based detection and correction of real-word errors.

Yarowsky (1994) experiments the use of decision lists for lexical ambiguity resolution, using context features (cf. section 1.4) like local syntactic patterns and collocational information, so that multiple types of evidence are considered in the context of an ambiguous word. In addition to word forms, the patterns involve part of speech tags and lemmas. The algorithm is evaluated in missing accent restoration task, in the case of restoring missing accents in Spanish and French text. It is evaluated against a predefined set of a few words giving an accuracy over 99%.

Golding and Schabes (1996) propose an hybrid method that combines part-of-speech trigrams and context features in order to detect and correct real-word errors. They present an experiment where their system has substantially higher performance than the grammar checker in Microsoft Word, but its coverage is limited to eighteen particular confusion sets composed by two or three similar words (e.g.: weather, whether).

The last three systems rely on a previously collected set of confusion sets (sets of similar words or accentuation ambiguities). On the contrary, our system has to choose a single proposal for any possible spelling error, and it is therefore impossible to collect the confusion sets (i.e. sets of proposals for each spelling error) beforehand. We also need to correct as many errors as possible, even if the amount of data for a particular case is scarce.

Conclusion

This work presents a study of different methods, which build on the correction proposals of *ispell*, aiming at giving a single correction proposal for misspellings. One of the difficult aspects of the problem is that of testing the results. For that reason, we used both a corpus with artificially generated errors for training and testing, and a corpus with genuine errors for testing.

Examining the results, we observe that the results improve as more context is taken into account. The word-form frequencies from the Brown Corpus serve as a crude but helpful criterion for choosing the correct proposal. The precision increases as closer contexts, like document frequencies, Constraint Grammar and context features are incorporated.

From the results on the corpus of genuine errors we can conclude the following. Firstly, the correct word is among *ispell*'s proposals 100% of the times, which means that all errors can be recovered. Secondly, the output that can be expected from our present system is that it will correct automatically the spelling errors with either 90% precision with full coverage and choosing a single proposal 24 times out of 25 (1.04 proposals left), or 93% precision with 90% coverage and leaving an average of 1.20 proposals.

Two of the techniques proposed, Brown Frequencies and Conceptual Density, did not yield useful results. CD only works for a very small fraction of the errors, which prevents us from making further conclusions.

There are reasons to expect better results in the future. First of all, the corpus with genuine errors contained very short documents, which caused the performance of DF to degrade substantially. Further tests with longer documents should yield better results. Secondly, we collected context features from an American English corpus which we used to correct British English texts. Once this language mismatch is solved better performance should be obtained. Lastly, there is room for improvement in the techniques themselves. We knowingly did not use any model of common misspellings. Regarding context features, only word-form features were collected, and part-of-speech and lemma features would presumably be a good complement. Although we would expect limited improvement, stronger methods to combine the techniques can also be tried.

Acknowledgements

This research was supported by the Basque Government, the University of the Basque Country and the CICYT (Comisión Interministerial de Ciencia y Tecnología).

References

Agirre E. and Rigau G. (1996) *Word sense disambiguation using conceptual density.* In Proceedings of COLING, Copenhagen, Denmark.

Golding A. and Schabes. Y. (1996) *Combining trigram-based and feature-based methods for context-sensitive spelling correction* . In Proceedings of the 34th Annual Meeting of the Association Computational Linguistics, Santa Cruz, CA.

Ispell (1993) International Ispell Version 3.1.00, 10/08/93.

Francis S., Kucera H. (1967) *Computing Analysis of Present-Day American English.* Brown University Press.

Karlsson F., Voutilainen A., Heikkila J. and Anttila A. (1995) *Constraint Grammar: a Language Independent System for Parsing Unrestricted Text.* Ed.Mouton de Gruyter .

Koskenniemi K. (1983) *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production.* Ph D. thesis, University of Helsinki.

Kukich K. (1992) *Techniques for automatically correcting words in text.* In ACM Computing Surveys, Vol. 24, N. 4, December, pp. 377-439.

Mays E., Damerau F. and Mercer. R. (1991) *Context based spelling correction* . Information Processing & Management, Vol. 27, N. 5, pp. 517-522.

Miller G. (1990) *Five papers on WordNet.* Special Issue of the International Journal of Lexicography, Vol. 3, N. 4.

Menezo J., Genthial D. and Courtin J. (1996) *Reconnaisances pluri-lexicales dans CELINE, un système multi-agents de détection et correction des erreurs.* NLP + IA 96, Moncton, N. B., Canada.

Yarowsky D. (1994) *Decision lists for lexical ambiguity resolution.* In Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, Las Cruces, NM, pp.88-95.

Disambiguating bilingual nominal entries against WordNet

German Rigau.[*]
Departament de Llenguatges i Sistemes Informˆtics. Universitat Polit   cnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. g.rigau@lsi.upc.es

Eneko Agirre.[**]
Lengoaia eta Sistema Informatikoak Saila. Euskal Herriko Unibertsitatea.
p.k. 649, 20080 Donostia. Spain. jibagbee@si.ehu.es

1. INTRODUCTION

One reason why the lexical capabilities of NLP systems have remained weak is because of the labour intensive nature of encoding lexical entries for the lexicon. It has been estimated that the average time needed to construct manually a lexical entry for a Machine Translation system is about 30 minutes [Neff et al. 93]. The automatic acquisition of lexical knowledge is the main field of the research work presented here. In particular, this paper explores the acquisition of conceptual knowledge from bilingual dictionaries (French/English, Spanish/English and English/Spanish) using a pre-existing broad coverage Lexical Knowledge Base (LKB) WordNet [Miller 90].

The automatic acquisition of lexical knowledge from monolingual machine-readable dictionaries (MRDs) has been broadly explored (e.g. [Boguraev & Briscoe 90], [Artola 93], [Castellón 93], [Wilks et al. 93], [Dolan et al. 93]), while less attention has been paid to bilingual dictionaries (e.g. [Ageno et al. 94], [Knight & Luk 94]).

Bilingual dictionaries contain information about the connection of vocabularies in two different languages. However, MRDs are made for human readers and the information contained in it is not immediately usable as a computational lexicon. For instance word translations are not marked with a sense or group of senses (sense mismatch problem), but they are sometimes annotated with subject field codes or cue words in the source language.

Two different, complementary approaches are explored in this paper. Both of them use WordNet to obtain a multilingual LKB (MLKB). The resulting MLKB has the same structure as WordNet, but some nodes are attached additionally to disambiguated vocabulary of other languages.

In one of the approaches each entry of the dictionary is taken in turn, exploiting the information in the entry itself. The inferential capability for disambiguating the translation is given by Semantic Density over WordNet [Agirre & Rigau, 95]. In the other approach, the bilingual dictionary was merged with WordNet, exploiting mainly synonymy relations. Each of the approaches was used in a different dictionary. The first approach was used on a French-English dictionary (using one direction only), and the second approach on a Spanish-English/English-Spanish dictionary (both directions).

[*] German Rigau was supported by a grant from the Ministerio de Educación y Ciencia.

[**] Eneko Agirre was supported by a grant from the Basque Government.

After this short introduction, section 2 shows some experiments and results using Semantic Density on the bilingual French/English dictionary. In section 3 several complementary techniques and results using the Spanish bilingual dictionaries are explained.

2. WORD SENSE DISAMBIGUATION USING CONCEPTUAL DENSITY

2.1 The French/English bilingual dictionary

The French/English bilingual dictionary contains 21,322 entries. Each entry can comprise several or a single sense of the source word, which in the scope of this paper we will call subentries. For instance, the entry for ‘maintien’ is split in two subentries:

```
maintien n.m. (attitude) bearing; (conservation) maintenance.

maintien 1: n.m. (attitude) bearing
maintien 2: n.m. (conservation) maintenance
```

The dictionary has 31,502 such subentries, from which 16,917 are nominal subentries.

Each subentry can have the following fields: part of speech (always), semantic field (one out of a set of 20, e.g. `comm.` in `tr sor 2` in the example below), cue in French (e.g. `ressources` in `tr sor 2`) and one or several translations in English (always). The semantic field and the cue in French are used to determine the context or the usage of the French word when translated by the subentry.

```
folie 1: n.f. madness
provision 1: n.f. supply, store
tr sor 2: n.m. (ressources) (comm.) finances
```

In order to figure out which WordNet sense(s) fit(s) best the French headword, the algorithm needs contextual information (as we humans do). If we do not have any contextual information, and the translation has more than one sense, it is not possible to find the correct sense(s)[1] . The cases where we can try to disambiguate the translation are the following:

1) one of the translation words is monosemous in WordNet
2) the translation is given by a list of words
3) a cue in French is provided alongside the translation
4) a semantic field is provided

From the examples above, `folie`’s translation has more than one sense and therefore is not a member of any of the cases. `provision` has two translation polysemous translations and therefore belongs to case 2. `tr sor` has a monosemous translation and also comes with a French cue (`ressources`) and a semantic field (`comm` meaning commercial), and therefore belongs to cases 2, 3 and 4.

The figures for combinations of the above cases found in the bilingual dictionaries are the following:

[1] In this work we try to assign a single sense to the translations.

translation not in WordNet	4,081	24%
unique translation, n senses	4,761	28%
any combination of cases 1,2,3,4	8,075	48%
total	16,917	100%

Table 1

The figures mean that, from all the senses of French nouns, we can disambiguate at most 48% of them. The coverage of WordNet is not very impressive, only 76% of the English nouns in the bilingual dictionary. This is caused by several problems that will be dealt with below.

The bilingual subentries that provide disambiguation information have the distribution shown below. Some subentries belong at the same time to more than one case.

case 1; 1 sense	5,039	30%
case 2; more than one translation	630	4%
case 3; cue in French	2,954	17%
case 4; semantic field	1,067	6%

Table 2

Those that have a monosemous unique translation can be directly linked. Besides we still have not experimented with the use of semantic fields. Therefore, the algorithm will focus on bilingual subentries with multiple translations and/or cues in French.

2.2 Treatment of complex translations and cues

In the previous paragraph, it was said that 24% of the translations were not found in WordNet. A quick look at some of the translations revealed that the failure was sometimes caused by the translation being in a plural form, being composed by a whole noun phrase, brackets, etc. The same situation was observed in the cues, which were often composed by a phrase or a list of phrases. We call these translations and cues *complex*. Some examples of complex translations and cues follow:

```
batterie 2: n.f. (mus.) drums
e'poux 2: n.m. the married couple
escale 2: n.f. (port) port of call
microplaquette 1: n.f. (micro) chip
remonte'e 2: n.f. (d'eau, de prix) rise
```

The treatment for the translations and cues that could not be found directly in WordNet or the bilingual dictionary respectively was done in two steps. First, a morphological analysis was performed, and if it was not successful, combinations of the component words were tried.

A) morphological analysis: For English we use the morphological analyser provided by WordNet. In the case of French, a naive morphological analysis is tried (valid for nouns only), checking the resulting potential lemmas against the bilingual dictionary itself. For instance, morphological lookup for the translation for `batterie 2` would yield `drum`.

B) complex phrases: when the translation or cue is composed by more than one word, several combinations of the component words are tried. The longest combination of words that is successfully looked-up is returned. If no combination is succesful, then all the component words that are correct nouns (according to WordNet for English, and the bilingual dictionary for French) are returned. For the translation of `e'poux 2` this procedure would return `married couple`, which is correctly found in WordNet. In another example, `port of call` would yield both `port` and `call`. The same applies for cues: the processing of the cue `d'eau, de prix` would output both `eau` and `prix`. Brackets are also taken into account, but in this case the words inside brackets would never be returned on their own, only as components of a compound noun.

A sample of 50 complex translations was evaluated, to see the reliability of the method proposed. In 21% of the results, the single correct translation was proposed. The most significant part of the translation was captured in 67% of the cases, and only 12% of the proposed translations were wrong.

After processing the English translations, it was found that the coverage of WordNet increased from 76% to 95%, leaving only 891 subentries that could not be processed. This means that the figures for all cases in tables 1 and 2 change, as shown in tables 1' and 2'.

translation not in WordNet	891	5%
unique translation, n senses	6,440	38%
any combination of cases 1,2,3,4	9,586	57%
total	16,917	100%

Table 1'

case 1; 1 sense	5,119	30%
case 2; more than one translation	958	6%
case 3; cue in French	3,702	22%
case 4; semantic field	1,365	8%

Table 2'

2.3 The disambiguation procedure

In the core of the disambiguation procedure we use conceptual density as described in [Agirre & Rigau, 95], [Rigau 94] and [Agirre et al. 94]. Conceptual Density provides a basis for determining relatedness among words, taking as reference a structured hierarchical net which in this case is WordNet. For instance, in figure 1 we have a word W with four senses. Each sense belongs to a subtree in the hierarchical net. The dots in the subtrees represent the senses of either the word to be disambiguated (W) or the words in the context. Semantic Density will yield the highest density for the subtree containing more senses of those, relative to the total amount of senses in the subtree.

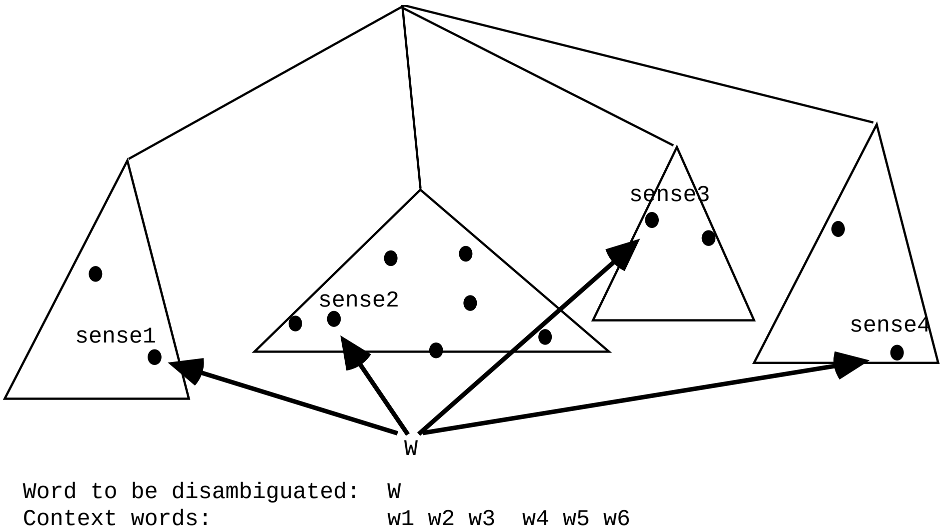


Figure 1: senses of a word in WordNet

The relatedness of a certain word-sense to the words in the context allows us to select that sense over the others. Following with the example in figure 1, sense2 would be chosen for W, because it belongs to the subtree with highest Semantic Density. In some cases more than one sense of the word to be disambiguated will belong to the selected subtree. In that case multiple senses are returned.

The context words are provided by the cue words in French and multiple translations. Cue words are in French, and therefore need to be translated into English, which is done using the bilingual dictionary.

In order to evaluate the contribution of each kind of contextual information separately, two experiments where performed on two sets of subentries: a set comprising French cues with a single translation word, and a set containing more than one translation but without any French cue.

2.4 Estimate the contribution of French cues

French cues are looked up in the bilingual dictionary, and all the English translations of the cue are input to the algorithm alongside the English translation. These English words will provide the necessary contextual information for the disambiguation of the translation.

A set of experiments was performed to evaluate the expected precision when disambiguating subentries that had a single English translation and a French cue. For this purpose, 59 French subentries fulfilling the given condition were selected at random

The precision and coverage are shown in the second line of the table below. The precision is considerably higher than random guessing[2]. The error rate was deemed too high, specially for some of the potential applications. In order to reduce the error rate several heuristics were tried. Declining to disambiguate translations with more than 5 senses was the most successful. As the third line of the following table shows, precision

[2] The figure for random guessig takes into account all noun entries. It was obtained analytically using the polysemy figures for all translations.

raised at the cost of the coverage.

	precision	coverage
random guessing	44.8%	-
original results	67.4%	72.9%
heuristic	83.3%	50.8%

Table 3

2.5 Estimate contribution of several translations

In this experiment 30 subentries that had more than one English translation were selected at random. The disambiguation algorithm was fed with the set of translation words and produced a set of WordNet synsets. The results, with and without applying the heuristic, are the following:

	precision	coverage
random guessing	44.8%	-
original results	89.3%	93.3%
heuristic	90.9%	73.3%

Table 4

Performance for this subset of the definitions is considerably better than for French cues. The heuristic does not yield significant improvement in precision, and the original results are preferred.

2.6 Overall results

Table 5 summarises the overall results. The algorithm was run over all the subentries, except those containing semantic fields. This means that in the best case, 8,221[3] subentries (53% of the total 15,552) could be linked. For a given subentry, whether it was monosemous or not was checked first. If not, disambiguation using multiple translations was tried, and last, cues in French were used. Monosemous translations account for most of the links made. The low coverage when disambiguating with French cues accounts for most of the failures to make links.

no result	8,311	53%
result obtained	7,241	47%
case 1; 1 sense	5,119	33%
case 2; >1 trans	723	5%
case 3; cue	1,399	9%
total	15,552	100%

Table 5

The links made, as calculated in the previous experiments, are highly reliable. The confidence for monosemous links (case 1) would be 100% if it not were because of complex translations, for which 88% of precision can be expected. For case 2, 93% of correct answers can be expected which descends to 83% for case 3 subentries.

[3] Calculated from tables 1’ and 2’, substracting the number of semantic fields from the overall combination of cases 1,2,3 and 4.

Overall coverage of this method will hopefully improve when semantic fields are taken into account.

3. MERGING LEXICAL KNOWLEDGE RESOURCES

Four experiments have been performed exploiting simple properties to attach Spanish nouns from the Spanish/English-English/Spanish bilingual dictionary to noun synsets in WordNet 1.5.

The nominal part of WordNet 1.5 has 60557 synsets and 87642 English nouns (76127 monosemous). The Spanish/English bilingual dictionary contains 12370 Spanish nouns and 11467 English nouns in 19443 connections among them. On the other hand, the English/Spanish bilingual dictionary is less informative than the other one containing only 10739 English nouns, 10549 Spanish nouns in 16324 connections.

Merging both dictionaries a list of equivalence pairs of nouns have been obtained. The combined dictionary contains 15848 English nouns, 14880 Spanish nouns and 28131 connections.

For instance, for the word "masa" in Spanish the following list of equivalence pairs can be obtained:

```
----------------------- English/Spanish
bulk masa
dough masa
mass masa
----------------------- Spanish/English
cake masa
crowd_of_people masa
dough masa
ground masa
mass masa
mortar masa
volume masa
```

From the combined dictionary, there are only 12665 English nouns placed in WordNet 1.5 which represents 19383 synsets. That is, the maximum coverage we can expect of WordNet1.5 using both bilingual Spanish/English dictionaries is 32%. In the next table the summarised amount of data is shown.

	English nouns	Spanish nouns	synsets	connections
WordNet1.5	87,642	-	60,557	107,424
Spanish/English	11,467	12,370	-	19,443
English/Spanish	10,739	10,549	-	16,324
Merged Bilingual	15,848	14,880	-	28,131
Maximum Coverage	12,665	13,208	19,383	24,613
of WordNet	14%	-	32%	-
of bilingual	80%	90%	-	87%

Table 6

The connection of Spanish nouns to Synsets in WordNet 1.5 has been performed in the following cases:

1) Those Spanish nouns translations of monosemous English nouns (one sense in WordNet). Considering for instance that the noun abduction has only one sense in WordNet1.5[4] :

Synonyms/Hypernyms (Ordered by Frequency) of noun abduction
1 sense of abduction

```
Sense 1
<abduction>
      => <capture, seizure>
          => <felony>
              => <crime, law-breaking>
                  => <evildoing, transgression>
                      => <wrongdoing, misconduct>
                          => <activity>
                              => <act, human action, human activity>
```

and there are two possible translations for abduction for Spanish

secuestro	<-->	abduction
rapto	<-->	abduction

the following attachment has been produced:

<abduction>   <-->   <secuestro, rapto>

Only 6616 English nouns from the equivalence pairs list are monosemous (42% of the total English nouns). Thus, this simple approach has produced 9057 connections among 7636 Spanish nouns and 5963 synsets of WordNet1.5 with a very high degree of confidence. The polysemous degree in this case is 1.19 synsets per Spanish noun with 1.52 Spanish nouns per synset. Next table shows the results following this process.

[4] In the following examples, brackets are used indicating synsets (concepts) and => means hyponym-of.

	English nouns	Spanish nouns	synsets	connec.	Poly.	Syn.
WordNet	87,642	-	60,557	107,424	1.2	1.8
Bilingual	15,848	14,880	-	28,131		
Maximum Coverage	12,665	13,208	19,383	24,613	1.9	1.3
Case 1	6,616	**7,636**	5,963	9,057	1.2	1.5
of WordNet	8%	-	10%	-		
of Bilingual	42%	51%	-	-		
of Maximum	52%	58%	30%	37%		
of total	58%	63%	37%	37%		
Total	11,470	12,039	15,897	24,535		

Table 7

2) Those Spanish nouns with only one translation (although, the translation could be polysemous). Consider for instance the only translation found into the merged dictionary for the Spanish noun *anfibio* :

amphibian	<-->	anfibio

This process has produced three possible connections for the English WordNet1.5 amphibian:

<amphibian, amphibious vehicle>	<-->	<anfibio>
<amphibian, amphibious aircraft>	<-->	<anfibio>
<amphibian>	<-->	<anfibio>
=> <vertebrate, craniate>		

There are 8524 Spanish nouns with only one translation. These Spanish nouns are equivalence candidates of 7507 English nouns but only 6066 of these are present in WordNet1.5. Thus, this approach has generated 14164 connections among 7000 Spanish nouns and 10674 synsets. The polysemous ratio is 2.02 synsets per Spanish noun and there are 1.33 Spanish word per synset. In the following table the results for this approach are shown.

	English nouns	Spanish nouns	synsets	connec.	Poly.	Syn.
WordNet	87,642	-	60,557	107,424	1.2	1.8
Bilingual	15,848	14,880	-	28,131		
Maximum Coverage	12,665	13,208	19,383	24,613	1.9	1.3
Case 2	6,066	7,000	**10,674**	**14,164**	2.0	1.3
of WordNet	7%	-	18%	-		
of Bilingual	38%	47%	-	-		
of Maximum	48%	53%	55%	58%		
of total	53%	58%	67%	58%		
Total	11,470	12,039	15,897	24,535		

Table 8

3) Those English nouns (although, the translation could be polysemous) with only one translation. Consider the unique translation of banishment for the nominal part of the bilingual dictionaries:

banishment <--> destierro

Thus, the Spanish noun *destierro* has been attached to both synsets of banishment in WordNet:

```
<banishment, ostracism>   <-->     <destierro>
     => <exclusion>
        => <situation, state of affairs>
           => <state>

<banishment, proscription>          <-->     <destierro>
     => <rejection>
        => <act, human action, human activity>
```

There are 10285 English nouns with only one translation (out of 7383 are present in WordNet). These English nouns are equivalence translations of 8556 Spanish nouns. In this case, 11089 connections have been produced among 6470 Spanish nouns and 10223 synsets. Thus, the polysemous ratio is 1.71 synsets per Spanish noun with 1.08 Spanish noun per synset. In next table this data is summarized.

	English nouns	Spanish nouns	synsets	connec.	Poly.	Syn.
WordNet	87,642	-	60,557	107,424	1.2	1.8
Bilingual	15,848	14,880	-	28,131		
Maximum Coverage	12,665	13,208	19,383	24,613	1.9	1.3
Case 3	**7,383**	6,470	10,223	11,089	1.7	1.1
of WordNet	8%	-	17%	-		
of Bilingual	47%	44%	-	-		
of Maximum	58%	49%	53%	45%		
of total	64%	54%	64%	45%		
Total	11,470	12,039	15,897	24,535		

Table 9

4) Those synsets with several English nouns with the same translation. Consider the following translations for the word *error* in the merged bilingual dictionary:

error	<-->	error
mistake	<-->	error

then this process can generate the following attachment:

```
<mistake, error, fault>      <-->     <error>
     => <failure>
        => <nonaccomplishment, nonachievement>
           => <act, human action, human activity>

<error, mistake>             <-->     <error>
     => <misstatement>
        => <statement>
           => <message, content, subject matter, substance>
              => <communication>
                 => <social relation>
                    => <relation>
                       => <abstraction>
```

In this case, 3164 connections among 2261 Spanish nouns and 2195 synsets have been found. That means a polysemous ratio of 1.40 synsets per Spanish noun and 1.44 Spanish nouns per synset. The next table summarises the last approach.

	English nouns	Spanish nouns	synsets	connec.	Poly.	Syn.
WordNet	87,642	-	60,557	107,424	1.2	1.8
Bilingual	15,848	14,880	-	28,131		
Maximum Coverage	12,665	13,208	19,383	24,613	1.9	1.3
Case 4	2,092	2,261	2,195	3,164	1.4	1.4
of WordNet	2%	-	4%	-		
of Bilingual	13%	15%	-	-		
of Maximum	17%	17%	11%	13%		
of total	18%	19%	14%	13%		
Total	11,470	12,039	15,897	24,535		

Table 10

Merging all the connections we have obtained a micro-Spanish WordNet (with errors). The resulting data has 24535 connections among 12039 Spanish nouns and 15897 synsets of WordNet1.5. That is to say, a polysemous ratio of 2.03 synsets per Spanish noun with 1.54 synonymy degree. The next table shows the overall data:

	English nouns	Spanish nouns	synsets	connec.	Poly.	Syn.
WordNet	87,642		60,557	107,424	1.2	1.8
Bilingual	15,848	14,880		28,131		
Maximum Coverage	12,665	13,208	19,383	24,613	1.9	1.3
Case 1	6,616	**7,636**	5,963	9,057	1.2	1.5
Case 2	6,066	7,000	**10,674**	**14,164**	2.0	1.3
Case 3	**7,383**	6,470	10,223	11,089	1.7	1.1
Case 4	2,092	2,261	2,195	3,164	1.4	1.4
Total	11,470	12,039	15,897	24,535	2.0	1.5
of WordNet	13%	-	26%	-		
of Bilingual	72%	80%	-	-		
of Maximum	90%	91%	82%	100%		

Table 11

We have tested manually one hundred connections. 78 out of 100 were correct. Obviously, the most productive cases are the cases that introduce more errors.

4. CONSIDERATIONS

This paper shows that disambiguating bilingual nominal entries, and therefore linking bilingual dictionaries to WordNet is a feasible task. The complementary approaches presented here, Semantic Density on entry information and merging taking profit of dictionary structure, both attain high levels of precision on their own. The combination of both techniques, alongside using the semantic fields left aside by the first approach, should yield better precision and a raise in coverage. For instance, the first approach focuses on the information in the French/English direction of the dictionary, without using the reverse direction or exploiting the structure of the dictionary as in the second approach. The second approach, on the other hand, could take profit from both the information in each entry and the inferential capability of Semantic Density.

REFERENCES

[Ageno et al. 94] Ageno A., Castell—n I., Ribas F., Rigau G., Rodr’guez H., Samiotou A., *TGE: Tlink Generation Environment.* In Proceedings of the 16th International Conference on Computational Linguistics (Coling'94). Kyoto, Japan.

[Agirre et al. 94] Agirre E., Arregi X., Artola X., D’az de Ilarraza A. and Sarasola K., *Conceptual Distance and Automatic Spelling Correction*, in Workshop on Computational Linguistics for Speech and Handwriting Recognition, Leeds, 1994.

[Agirre & Rigau 95] Agirre E., Rigau G. *A Proposal for Word Sense Disambiguation using conceptual Distance*, submitted to the International Conference on Recent Advances in Natural Language Processing. Velingrad, Bulgaria. September 1995

[Artola 93] Artola X. *Conception et construccion d’un systeme intelligent d’aide diccionariale (SIAD)*. Ph. Thesis, Euskal Herriko Unibertsitatea, Donostia, 1993.

[Boguraev & Briscoe 90] Boguraev B. and Briscoe T. editors, *Computational Lexicography for Natural Language Processing.* Longman, Cambridge, England. 1990.

[Castellón 93] Castell—n I., *Lexicografia Computacional: Adquisici—n Autom‡tica de Conocimiento L xico,* Ph. Thesis, Universitat de Barcelona, Barcelona, 1993.

[Dolan et al. 93] Dolan W., Vanderwende L. and Richard son S., *Automatically deriving structured knowledge bases from on-line dictionaries.* in proceedings of the first Conference of the Pacific Association for Computational Linguistics (Pacling’93), April 21-24, Simon Fraser University, Vancouver, Canada. 1993.

[Knight & Luk 94] Knight K. and Luk S., *Building a Large-Scale Knowledge Base for Machine Translation,* in proceedings of the American Association for Artificial Inteligence. 1994.

[Miller 90] Miller G., *Five papers on WordNet,* Special Issue of International Journal of Lexicogrphy 3(4). 1990.

[Neff et al. 93] Neff M., Blaser B., Lange J-M., Lehmann H. and Dominguez. *Get it where you can: Acquiring and Maintaining Bilingual Lexicons for Machine Translation*, Paper presented at the AAAI Spring Simposium on Building Lexicons for Machine Translation, Stanford University.

[Rigau 94] Rigau G., *An Experiment on Automatic Semantic Tagging of Dictionary Senses,* in Proceedings of the International Workshop The Future of the Dictionary, Uriage-les-Bains, Grenoble, France, 1994.

[Wilks et al. 93] Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B., *Providing Machine Tractablle Dictionary Tools,* in Semantics and the Lexicon (Pustejowsky J. ed.), 341-401, 1993.

Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation *

German Rigau, Jordi Atserias
Dept. de Llenguatges i Sist. Informàtics
Universitat Politècnica de Catalunya
Barcelona, Catalonia
{g.rigau,batalla}@lsi.upc.es

Eneko Agirre
Lengoaia eta Sist. Informatikoak saila
Euskal Herriko Unibertsitatea
Donostia, Basque Country
jibagbee@si.ehu.es

Abstract

This paper presents a method to combine a set of unsupervised algorithms that can accurately disambiguate word senses in a large, completely untagged corpus. Although most of the techniques for word sense resolution have been presented as stand-alone, it is our belief that full-fledged lexical ambiguity resolution should combine several information sources and techniques. The set of techniques have been applied in a combined way to disambiguate the genus terms of two machine-readable dictionaries (MRD), enabling us to construct complete taxonomies for Spanish and French. Tested accuracy is above 80% overall and 95% for two-way ambiguous genus terms, showing that taxonomy building is not limited to structured dictionaries such as LDOCE.

1 Introduction

While in English the "lexical bottleneck" problem (Briscoe, 1991) seems to be softened (e.g. WordNet (Miller, 1990), Alvey Lexicon (Grover et al., 1993), COMLEX (Grishman et al., 1994), etc.) there are no available wide range lexicons for natural language processing (NLP) for other languages. Manual construction of lexicons is the most reliable technique for obtaining structured lexicons but is costly and highly time-consuming. This is the reason for many researchers having focused on the massive acquisition of lexical knowledge and semantic information from pre-existing structured lexical resources as automatically as possible.

As dictionaries are special texts whose subject matter is a language (or a pair of languages in the case of bilingual dictionaries) they provide a wide range of information about words by giving definitions of senses of words, and, doing that, supplying knowledge not just about language, but about the world itself.

One of the most important relation to be extracted from machine-readable dictionaries (MRD) is the hyponym/hypernym relation among dictionary senses (e.g. (Amsler, 1981), (Vossen and Serail, 1990)) not only because of its own importance as the backbone of taxonomies, but also because this relation acts as the support of main inheritance mechanisms helping, thus, the acquisition of other relations and semantic features (Cohen and Loiselle, 1988), providing formal structure and avoiding redundancy in the lexicon (Briscoe et al., 1990). For instance, following the natural chain of dictionary senses described in the *Diccionario General Ilustrado de la Lengua Española* (DGILE, 1987) we can discover that a *bonsai* is a cultivated plant or bush.

bonsai_1_2 *planta y arbusto así cultivado.*
(bonsai, plant and bush cultivated in that way)

The hyponym/hypernym relation appears between the entry word (e.g. *bonsai*) and the genus term, or the core of the phrase (e.g. *planta* and *arbusto*). Thus, usually a dictionary definition is written to employ a genus term combined with differentia which distinguishes the word being defined from other words with the same genus term[1].

As lexical ambiguity pervades language in texts, the words used in dictionary are themselves lexically ambiguous. Thus, when constructing complete disambiguated taxonomies, the correct dictionary sense of the genus term must be selected in each dictionary

*This research has been partially funded by CICYT TIC96-1243-C03-02 (ITEM project) and the European Comission LE-4003 (EuroWordNet project).

[1] For other kind of definition patterns not based on genus, a genus-like term was added after studying those patterns.

	DGILE		LPPL	
	overall	nouns	overall	nouns
headwords	93,484	53,799	15,953	10,506
senses	168,779	93,275	22,899	13,740
total number of words	1,227,380	903,163	97,778	66,323
average length of definition	7.26	9.68	3.27	3.82

Table 1: Dictionary Data

definition, performing what is usually called Word Sense Disambiguation (WSD)[2]. In the previous example *planta* has thirteen senses and *arbusto* only one.

Although a large set of dictionaries have been exploited as lexical resources, the most widely used monolingual MRD for NLP is LDOCE which was designed for learners of English. It is clear that different dictionaries do not contain the same explicit information. The information placed in LDOCE has allowed to extract other implicit information easily, e.g. taxonomies (Bruce et al., 1992). Does it mean that only highly structured dictionaries like LDOCE are suitable to be exploited to provide lexical resources for NLP systems?

We explored this question probing two disparate dictionaries: *Diccionario General Ilustrado de la Lengua Española* (DGILE, 1987) for Spanish, and *Le Plus Petit Larousse* (LPPL, 1980) for French. Both are substantially poorer in coded information than LDOCE (LDOCE, 1987)[3]. These dictionaries are very different in number of headwords, polysemy degree, size and length of definitions (c.f. table 1). While DGILE is a good example of a large sized dictionary, LPPL shows to what extent the smallest dictionary is useful.

Even if most of the techniques for WSD are presented as stand-alone, it is our belief, following the ideas of (McRoy, 1992), that full-fledged lexical ambiguity resolution should combine several information sources and techniques. This work does not address all the heuristics cited in her paper, but profits from techniques that were at hand, without any claim of them being complete. In fact we use unsupervised techniques, i.e. those that do not require hand-coding of any kind, that draw knowledge from a variety of sources – the source dictionaries, bilingual dictionaries and WordNet – in diverse ways. This paper tries to proof that using an appropriate method to combine those heuristics we can disambiguate the genus terms with reasonable precision, and thus construct complete taxonomies from any conventional dictionary in any language.

This paper is organized as follows. After this short introduction, section 2 shows the methods we have applied. Section 3 describes the test sets and shows the results. Section 4 explains the construction of the lexical knowledge resources used. Section 5 discusses previous work, and finally, section 6 faces some conclusions and comments on future work.

2 Heuristics for Genus Sense Disambiguation

As the methods described in this paper have been developed for being applied in a combined way, each one must be seen as a container of some part of the knowledge (or heuristic) needed to disambiguate the correct hypernym sense. Not all the heuristics are suitable to be applied to all definitions. For combining the heuristics, each heuristic assigns each candidate hypernym sense a normalized weight, i.e. a real number ranging from 0 to 1 (after a scaling process, where maximum score is assigned 1, c.f. section 2.9). The heuristics applied range from the simplest (e.g. heuristic 1, 2, 3 and 4) to the most informed ones (e.g. heuristics 5, 6, 7 and 8), and use information present in the entries under study (e.g. heuristics 1, 2, 3 and 4) or extracted from the whole dictionary as a unique lexical knowledge resource (e.g. heuristics 5 and 6) or combining lexical knowledge from several heterogeneous lexical resources (e.g. heuristic 7 and 8).

2.1 Heuristic 1: Monosemous Genus Term

This heuristic is applied when the genus term is monosemous. As there is only one hypernym sense candidate, the hyponym sense is attached to it. Only 12% of noun dictionary senses have monosemous genus terms in DGILE, whereas the smaller LPPL reaches 40%.

2.2 Heuristic 2: Entry Sense Ordering

This heuristic assumes that senses are ordered in an entry by frequency of usage. That is, the most used and important senses are placed in the entry before less frequent or less important ones. This heuristic provides the maximum score to the first sense of the hypernym candidates and decreasing scores to the others.

[2] Called also Lexical Ambiguity Resolution, Word Sense Discrimination, Word Sense Selection or Word Sense Identification.

[3] In LDOCE, dictionary senses are explicitly ordered by frequency, 86% dictionary senses have semantic codes and 44% of dictionary senses have pragmatic codes.

2.3 Heuristic 3: Explicit Semantic Domain

This heuristic assigns the maximum score to the hypernym sense which has the same semantic domain tag as the hyponym. This heuristic is of limited application: LPPL lacks semantic tags, and less than 10% of the definitions in DGILE are marked with one of the 96 different semantic domain tags (e.g. *med.* for medicine, or *der.* for law, etc.).

2.4 Heuristic 4: Word Matching

This heuristic trusts that related concepts will be expressed using the same content words. Given two definitions – that of the hyponym and that of one candidate hypernym – this heuristic computes the total amount of content words shared (including headwords). Due to the morphological productivity of Spanish and French, we have considered different variants of this heuristic. For LPPL the match among lemmas proved most useful, while DGILE yielded better results when matching the first four characters of words.

2.5 Heuristic 5: Simple Cooccurrence

This heuristic uses cooccurrence data collected from the whole dictionary (see section 4.1 for more details). Thus, given a hyponym definition (O) and a set of candidate hypernym definitions, this method selects the candidate hypernym definition (E) which returns the maximum score given by formula (1):

$$SC(O,E) = \sum_{w_i \in O \wedge w_j \in E} cw(w_i, w_j) \qquad (1)$$

The cooccurrence weight (cw) between two words can be given by Cooccurrence Frequency, Mutual Information (Church and Hanks, 1990) or Association Ratio (Resnik, 1992). We tested them using different context window sizes. Best results were obtained in both dictionaries using the Association Ratio. In DGILE window size 7 proved the most suitable, whereas in LPPL whole definitions were used.

2.6 Heuristic 6: Cooccurrence Vectors

This heuristic is based on the method presented in (Wilks et al., 1993) which also uses cooccurrence data collected from the whole dictionary (c.f. section 4.1). Given a hyponym definition (O) and a set of candidate hypernym definitions, this method selects the candidate hypernym (E) which returns the maximum score following formula (2):

$$CV(O,E) = sim(V_O, V_E) \qquad (2)$$

The similarity (sim) between two definitions can be measured by the dot product, the cosine function or the Euclidean distance between two vectors (V_O and V_E) which represent the contexts of the words presented in the respective definitions following formula (3):

$$V_{Def} = \sum_{w_i \in Def} civ(w_i) \qquad (3)$$

The vector for a definition (V_{Def}) is computed adding the cooccurrence information vectors of the words in the definition ($civ(w_i)$). The cooccurrence information vector for a word is collected from the whole dictionary using Cooccurrence Frequency, Mutual Information or Association Ratio. The best combination for each dictionary vary: whereas the dot product, Association Ratio, and window size 7 proved best for DGILE, the cosine, Mutual Information and whole definitions were preferred for LPPL.

2.7 Heuristic 7: Semantic Vectors

Because both LPPL and DGILE are poorly semantically coded we decided to enrich the dictionary assigning automatically a semantic tag to each dictionary sense (see section 4.2 for more details). Instead of assigning only one tag we can attach to each dictionary sense a vector with weights for each of the 25 semantic tags we considered (which correspond to the 25 lexicographer files of WordNet (Miller, 1990)). In this case, given an hyponym (O) and a set of possible hypernyms we select the candidate hypernym (E) which yields maximum similarity among semantic vectors:

$$SV(O,E) = sim(V_O, V_E) \qquad (4)$$

where sim can be the dot product, cosine or Euclidean Distance, as before. Each dictionary sense has been semantically tagged with a vector of semantic weights following formula (5).

$$V_{Def} = \sum_{w_i \in Def} swv(w_i) \qquad (5)$$

The salient word vector (swv) for a word contains a saliency weight (Yarowsky, 1992) for each of the 25 semantic tags of WordNet. Again, the best method differs from one dictionary to the other: each one prefers the method used in the previous section.

2.8 Heuristic 8: Conceptual Distance

Conceptual distance provides a basis for determining closeness in meaning among words, taking as reference a structured hierarchical net. Conceptual distance between two concepts is essentially the length of the shortest path that connects the concepts in the hierarchy. In order to apply conceptual distance, WordNet was chosen as the hierarchical knowledge base, and bilingual dictionaries were used to link Spanish and French words to the English concepts.

Given a hyponym definition (O) and a set of candidate hypernym definitions, this heuristic chooses the hypernym definition (E) which is closest according to the following formula:

$$CD(O,E) = dist(headword_O, genus_E) \tag{6}$$

That is, Conceptual Distance is measured between the headword of the hyponym definition and the genus of the candidate hypernym definitions using formula (7), c.f. (Agirre et al., 1994). To compute the distance between any two words (w_1,w_2), all the corresponding concepts in WordNet (c_{1_i}, c_{2_j}) are searched via a bilingual dictionary, and the minimum of the summatory for each concept in the path between each possible combination of $c_{1_i}$and c_{2_j} is returned, as shown below:

$$dist(w_1, w_2) = \min_{\substack{c_{1_i} \in w_1 \\ c_{2_j} \in w_2}} \sum_{\substack{c_k \in \\ path(c_{1_i}, c_{2_j})}} \frac{1}{depth(c_k)} \tag{7}$$

Formulas (6) and (7) proved the most suitable of several other possibilities for this task, including those which included full definitions in (6) or those using other Conceptual Distance formulas, c.f. (Agirre and Rigau, 1996).

2.9 Combining the heuristics: Summing

As outlined in the beginning of this section, the way to combine all the heuristics in one single decision is simple. The weights each heuristic assigns to the rivaling senses of one genus are normalized to the interval between 1 (best weight) and 0. Formula (8) shows the normalized value a given heuristic will give to sense E of the genus, according to the weight assigned to the heuristic to sense E and the maximum weight of all the sense of the genus E_i.

$$vote(O,E) = \frac{weight(O,E)}{\max_{E_i}(weigth(O,E_i))} \tag{8}$$

The values thus collected from each heuristic, are added up for each competing sense. The order in which the heuristics are applied has no relevance at all.

	DGILE	LPPL
Test Sampling	391	115
Correct Genus Selected	382 (98%)	111 (97%)
Monosemous	61 (16%)	40 (36%)
Senses per genus	2.75	2.29
idem (polysemous only)	3.64	3.02
Correct senses per genus	1.38	1.05
idem (polysemous only)	1.51	1.06

Table 2: Test Sets

3 Evaluation

3.1 Test Set

In order to test the performance of each heuristic and their combination, we selected two test sets at random (one per dictionary): 391 noun senses for DGILE and 115 noun senses for LPPL, which give confidence rates of 95% and 91% respectively. From these samples, we retained only those for which the automatic selection process selected the correct genus (more than 97% in both dictionaries). Both test sets were disambiguated by hand. Where necessary multiple correct senses were allowed in both dictionaries. Table 2 shows the data for the test sets.

3.2 Results

Table 3 summarizes the results for polysemous genus.

In general, the results obtained for each heuristic seem to be poor, but always over the random choice baseline (also shown in tables 3 and 4). The best heuristics according to the recall in both dictionaries is the sense ordering heuristic (2). For the rest, the difference in size of the dictionaries could explain the reason why cooccurrence-based heuristics (5 and 6) are the best for DGILE, and the worst for LPPL. Semantic distance gives the best precision for LPPL, but chooses an average of 1.25 senses for each genus.

With the combination of the heuristics (Sum) we obtained an improvement over sense ordering (heuristic 2) of 9% (from 70% to 79%) in DGILE, and of 7% (from 66% to 73%) in LPPL, maintaining in both cases a coverage of 100%. Including monosemous genus in the results (c.f. table 4), the sum is able to correctly disambiguate 83% of the genus in DGILE (8% improvement over sense ordering) and 82% of the genus in LPPL (4% improvement). Note that we are adding the results of eight different heuristics with eight different performances, improving the individual performance of each one.

In order to test the contribution of each heuristic to the total knowledge, we tested the sum of all the heuristics, eliminating one of them in turn. The results are provided in table 5.

LPPL	random	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Sum
recall	36%	-	66%	-	8%	11%	22%	11%	50%	73%
precision	36%	-	66%	-	66%	44%	61%	57%	76%	73%
coverage	100%	-	100%	-	12%	25%	36%	19%	66%	100%
DGILE										
recall	30%	-	70%	1%	44%	57%	60%	57%	47%	79%
precision	30%	-	70%	100%	72%	57%	60%	58%	49%	79%
coverage	100%	-	100%	1%	61%	100%	100%	99%	95%	100%

Table 3: Results for polysemous genus.

LPPL	random	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Sum
recall	59%	35%	78%	-	40%	42%	50%	42%	68%	82%
precision	59%	100%	78%	-	93%	82%	84%	88%	87%	82%
coverage	100%	35%	100%	-	43%	51%	59%	48%	78%	100%
DGILE										
recall	41%	16%	75%	2%	41%	59%	63%	59%	48%	83%
precision	41%	100%	75%	100%	79%	65%	66%	63%	57%	83%
coverage	100%	16%	100%	2%	56%	95%	97%	94%	89%	100%

Table 4: Overall results.

LPPL	Sum	-(1)	-(2)	-(3)	-(4)	-(5)	-(6)	-(7)	-(8)
recall	82%	73%	74%	-	73%	76%	77%	77%	78%
precision	82%	73%	75%	-	73%	76%	77%	77%	78%
coverage	100%	100%	99%	-	100%	100%	100%	100%	100%
DGILE									
recall	83%	79%	72%	81%	81%	81%	81%	81%	77%
precision	83%	79%	72%	82%	81%	81%	81%	81%	77%
coverage	100%	100%	100%	98%	100%	100%	100%	100%	100%

Table 5: Knowledge provided by each heuristic (overall results).

(Gale et al., 1993) estimate that any sense-identification system that does not give the correct sense of polysemous words more than 75% of the time would not be worth serious consideration. As table 5 shows this is not the case in our system. For instance, in DGILE heuristic 8 has the worst performance (see table 4, precision 57%), but it has the second larger contribution (see table 5, precision decreases from 83% to 77%). That is, even those heuristics with poor performance can contribute with knowledge that other heuristics do not provide.

3.3 Evaluation

The difference in performance between the two dictionaries show that quality and size of resources is a key issue. Apparently the task of disambiguating LPPL seems easier: less polysemy, more monosemous genus and high precision of the sense ordering heuristic. However, the heuristics that depend only on the size of the data (5, 6) perform poorly on LPPL, while they are powerful methods for DGILE.

The results show that the combination of heuristics is useful, even if the performance of some of the heuristics is low. The combination performs better than isolated heuristics, and allows to disambiguate all the genus of the test set with a success rate of 83% in DGILE and 82% in LPPL.

All the heuristics except heuristic 3 can readily be applied to any other dictionary. Minimal parameter adjustment (window size, cooccurrence weigth formula and vector similarity function) should be done to fit the characteristics of the dictionary, but according to our results it does not alter significantly the results after combining the heuristics.

4 Derived Lexical Knowledge Resources

4.1 Cooccurrence Data

Following (Wilks et al., 1993) two words cooccur if they appear in the same definition (word order in definitions are not taken into account). For instance, for DGILE, a lexicon of 300,062 cooccurrence pairs among 40,193 word forms was derived (stop words were not taken into account). Table 6 shows the first eleven words out of the 360 which cooccur with *vino* (wine) ordered by Association Ratio. From left to right, Association Ratio and number of occurrences.

The lexicon (or machine-tractable dictionary,

AR	#oc.	
11.1655	15	*tinto* (red)
10.0162	23	*beber* (to drink)
9.6627	14	*mosto* (must)
8.6633	9	*jerez* (sherry)
8.1051	9	*cubas* (cask, barrel)
8.0551	16	*licor* (liquor)
7.2127	17	*bebida* (drink)
6.9338	12	*uva* (grape)
6.8436	9	*trago* (drink, swig)
6.6221	12	*sabor* (taste)
6.4506	15	*pan* (bread)

Table 6: Example of association ratio for *vino* (wine).

MTD) thus produced from the dictionary is used by heuristics 5 and 6.

4.2 Multilingual Data

Heuristics 7 and 8 need external knowledge, not present in the dictionaries themselves. This knowledge is composed of semantic field tags and hierarchical structures, and both were extracted from WordNet. In order to do this, the gap between our working languages and English was filled with two bilingual dictionaries. For this purpose, we derived a list of links for each word in Spanish and French as follows.

Firstly, each Spanish or French word was looked up in the bilingual dictionary, and its English translation was found. For each translation WordNet yielded its senses, in the form of WordNet concepts (synsets). The pair made of the original word and each of the concepts linked to it, was included in a file, thus producing a MTD with links between Spanish or French words and WordNet concepts. Obviously some of this links are not correct, as the translation in the bilingual dictionary may not necessarily be understood in its senses (as listed in WordNet). The heuristics using these MTDs are aware of this.

For instance when accessing the semantic fields for *vin* (French) we get a unique translation, wine, which has two senses in WordNet: `<wine,vino>` as a beverage, and `<wine, wine-coloured>` as a kind of color. In this example two links would be produced (*vin*, `<wine,vino>`) and (*vin*, `<wine, wine-coloured>`). This link allows us to get two possible semantic fields for *vin* (`noun.food`, file 13, and `noun.attribute`, file 7) and the whole structure of the hierarchy in WordNet for each of the concepts.

5 Comparison with Previous Work

Several approaches have been proposed for attaching the correct sense (from a set of prescribed ones) of a word in context. Some of them have been fully tested in real size texts (e.g. statistical methods (Yarowsky, 1992), (Yarowsky, 1994), (Miller and Teibel, 1991), knowledge based methods (Sussna, 1993), (Agirre and Rigau, 1996), or mixed methods (Richardson et al., 1994), (Resnik, 1995)). The performance of WSD is reaching a high stance, although usually only small sets of words with clear sense distinctions are selected for disambiguation (e.g. (Yarowsky, 1995) reports a success rate of 96% disambiguating twelve words with two clear sense distinctions each one).

This paper has presented a general technique for WSD which is a combination of statistical and knowledge based methods, and which has been applied to disambiguate all the genus terms in two dictionaries.

Although this latter task could be seen easier than general WSD[4], genus are usually frequent and general words with high ambiguity[5]. While the average of senses per noun in DGILE is 1.8 the average of senses per noun genus is 2.75 (1.30 and 2.29 respectively for LPPL). Furthermore, it is not possible to apply the powerful "one sense per discourse" property (Yarowsky, 1995) because there is no discourse in dictionaries.

WSD is a very difficult task even for humans[6], but semiautomatic techniques to disambiguate genus have been broadly used (Amsler, 1981) (Vossen and Serail, 1990) (Ageno et al., 1992) (Artola, 1993) and some attempts to do automatic genus disambiguation have been performed using the semantic codes of the dictionary (Bruce et al., 1992) or using cooccurrence data extracted from the dictionary itself (Wilks et al., 1993).

Selecting the correct sense for LDOCE genus terms, (Bruce et al., 1992) report a success rate of 80% (90% after hand coding of ten genus). This impressive rate is achieved using the intrinsic char-

[4] In contrast to other sense distinctions Dictionary word senses frequently differ in subtle distinctions (only some of which have to do with meaning (Gale et al., 1993)) producing a large set of closely related dictionary senses (Jacobs, 1991).

[5] However, in dictionary definitions the headword and the genus term have to be the same part of speech.

[6] (Wilks et al., 1993) disambiguating 197 occurrences of the word bank in LDOCE say "was not an easy task, as some of the usages of bank did not seem to fit any of the definitions very well". Also (Miller et al., 1994) tagging semantically SemCor by hand, measure an error rate around 10% for polysemous words.

acteristics of LDOCE. Furthermore, using only the implicit information contained into the dictionary definitions of LDOCE (Cowie et al., 1992) report a success rate of 47% at a sense level. (Wilks et al., 1993) reports a success rate of 45% disambiguating the word bank (thirteen senses LDOCE) using a technique similar to heuristic 6. In our case, combining informed heuristics and without explicit semantic tags, the success rates are 83% and 82% overall, and 95% and 75% for two-way ambiguous genus (DGILE and LPPL data, respectively). Moreover, 93% and 92% of times the real solution is between the first and second proposed solution.

6 Conclusion and Future Work

The results show that computer aided construction of taxonomies using lexical resources is not limited to highly-structured dictionaries as LDOCE, but has been succesfully achieved with two very different dictionaries. All the heuristics used are unsupervised, in the sense that they do not need hand-codding of any kind, and the proposed method can be adapted to any dictionary with minimal parameter setting.

Nevertheless, quality and size of the lexical knowledge resources are important. As the results for LPPL show, small dictionaries with short definitions can not profit from raw corpus techniques (heuristics 5, 6), and consequently the improvement of precision over the random baseline or first-sense heuristic is lower than in DGILE.

We have also shown that such a simple technique as just summing is a useful way to combine knowledge from several unsupervised WSD methods, allowing to raise the performance of each one in isolation (coverage and/or precision). Furthermore, even those heuristics with apparently poor results provide knowledge to the final result not provided by the rest of heuristics. Thus, adding new heuristics with different methodologies and different knowledge (e.g. from corpora) as they become available will certainly improve the results.

Needless to say, several improvements can be done both in individual heuristic and also in the method to combine them. For instance, the cooccurrence heuristics have been applied quite indiscriminately, even in low frequency conditions. Significance tests or association coefficients could be used in order to discard low confidence decisions. Also, instead of just summing, more clever combinations can be tried, such as training classifiers which use the heuristics as predictor variables.

Although we used these techniques for genus disambiguation we expect similar results (or even better taken the "one sense per discourse" property and lexical knowledge acquired from corpora) for the WSD problem.

7 Acknowledgments

This work would not be possible without the collaboration of our colleagues, specially Jose Mari Arriola, Xabier Artola, Arantza Diaz de Ilarraza, Kepa Sarasola and Aitor Soroa in the Basque Country and Horacio Rodríguez in Catalonia.

References

Alicia Ageno, Irene Castellón, Maria Antonia Martí, Francesc Ribas, German Rigau, Horacio Rodríguez, Mariona Taulé and Felisa Verdejo. 1992. SEISD: An environment for extraction of Semantic information from on-line dictionaries. In *Proceedings of the 3th Conference on Applied Natural Language Processing (ANLP'92)*, Trento, Italy.

Eneko Agirre, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza and Kepa Sarasola. 1994. Conceptual Distance and Automatic Spelling Correction. In *Proceedings of the workshop on Computational Linguistics for Speech and Handwriting Recognition*, Leeds, United Kingdom.

Eneko Agirre and German Rigau. 1996. Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, pages 16–22. Copenhagen, Denmark.

Robert Amsler. 1981. A Taxonomy for English Nouns and Verbs. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*, pages 133–138. Stanford, California.

Xabier Artola. 1993. *Conception et construccion d'un systeme intelligent d'aide diccionariale (SIAD)*. PhD. Thesis, Euskal Herriko Unibertsitatea, Donostia, Basque Country.

Eduard Briscoe, Ann Copestake and Branimir Boguraev. 1990. Enjoy the paper: Lexical Semantics via lexicology. In *Proceedings of the 13th International Conference on Computational Linguistics (Coling'90)*, pages 42–47.

Eduard Briscoe. 1991. Lexical Issues in Natural Language Processing. In Klein E. and Veltman F. eds. *Natural Language and Speech.* pages 39–68, Springer-Verlag.

Rebecca Bruce, Yorick Wilks, Louise Guthrie, Brian Slator and Ted Dunning. 1992. NounSense - A Disambiguated Noun Taxonomy with a Sense of

Humour. *Research Report MCCS-92-246.* Computing Research Laboratory, New Mexico State University. Las Cruces.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, vol. 16, ns. 1, 22-29.

P. Cohen and C. Loiselle. 1988. Beyond ISA: Structures for Plausible Inference in Semantic Data. In *Proceedings of 7th Natural Language Conference AAAI'88.*

Jim Cowie, Joe Guthrie and Louise Guthrie. 1992. Lexical Disambiguation using Simulated Annealing. In *Proceedings of DARPA WorkShop on Speech and Natural Language*, pages 238-242, New York.

DGILE 1987. *Diccionario General Ilustrado de la Lengua Española VOX.* Alvar M. ed. Biblograf S.A. Barcelona, Spain.

William Gale, Kenneth Church and David Yarowsky. 1993. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities* 26, pages 415-439.

Ralph Grishman, Catherine Macleod and Adam Meyers. 1994.. Comlex syntax: building a computational lexicon. In *Proceedings of the 15th Annual Meeting of the Association for Computational Linguistics, (Coling'94).* 268-272. Kyoto, Japan.

Claire Grover, John Carroll and John Reckers. 1993. The Alvey Natural Language Tools grammar (4th realese). *Technical Report 284.* Computer Laboratory, Cambridge University, UK.

Paul Jacobs. 1991. Making Sense of Lexical Acquisition. In Zernik U. ed., *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey.

LDOCE 1987. *Longman Dictionary of Contemporary English.* Procter, P. ed. Longman, Harlow and London.

LPPL 1980. *Le Plus Petit Larousse.* Gougenheim, G. ed. Librairie Larousse.

Sussan McRoy. 1992. Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics* 18(1).

George Miller. 1990. Five papers on WordNet. *Special Issue of International Journal of Lexicography* 3(4).

George Miller and David Teibel. 1991. A proposal for Lexical Disambiguation. In *Proceedings of DARPA Speech and Natural Language Workshop*, 395-399, Pacific Grave, California.

George Miller, Martin Chodorow, Shari Landes, Claudia Leacock and Robert Thomas. 1994. Using a Semantic Concordance for sense Identification. In *Proceedings of ARPA Workshop on Human Language Technology.*

Philip Resnik. 1992. WordNet and Distributional analysis: A class-based approach to lexical discovery. In *Proceedings of AAAI Symposyum on Probabilistic Approaches to NL*, San Jose, California.

Philip Resnik. 1995. Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of the Third Workshop on Very Large Corpora*, MIT.

R. Richardson, A.F. Smeaton and J. Murphy. 1994. Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words. *Working Paper CA-1294*, School of Computer Applications, Dublin City University. Dublin, Ireland.

Michael Sussna. 1993. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. In *Proceedings of the Second International Conference on Information and knowledge Management.* Arlington, Virginia.

Piek Vossen and Iskander Serail. 1992. Word-Devil, a Taxonomy-Browser for Lexical Decomposition via the Lexicon. *Esprit BRA-3030 Acquilex Working Paper n. 009.*

Yorick Wilks, Dam Fass, Cheng-Ming Guo, James McDonald, Tony Plate and Brian Slator. 1993. Providing Machine Tractable Dictionary Tools. In Pustejowsky J. ed. *Semantics and the Lexicon*, pages 341–401.

David Yarowsky. 1992. Word-Sense Disambiguation Using Statistical Models of Rogets Categories Trained on Large Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (Coling'92)*, pages 454-460. Nantes, France.

David Yarowsky. 1994. Decision Lists for Lexical Ambiguity Resolution. In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, (ACL'94).* Las Cruces, New Mexico.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics, (ACL'95).* Cambridge, Massachussets.