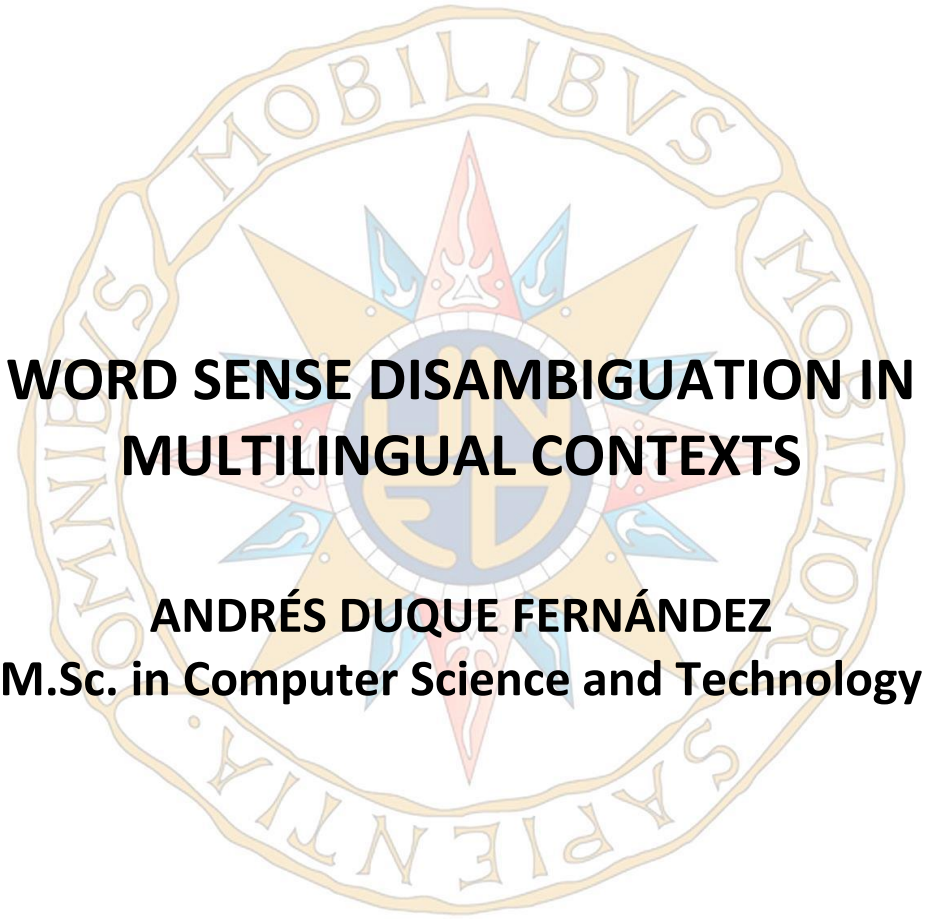
TESIS DOCTORAL

2017



WORD SENSE DISAMBIGUATION IN MULTILINGUAL CONTEXTS

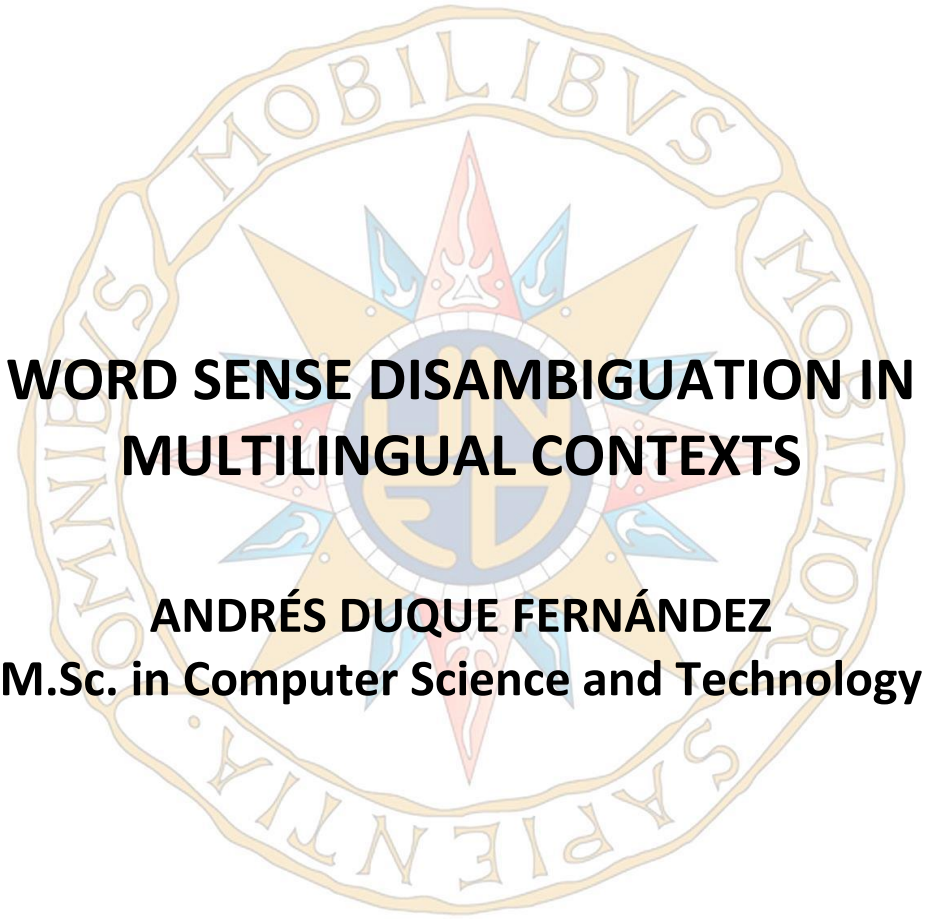
ANDRÉS DUQUE FERNÁNDEZ
M.Sc. in Computer Science and Technology

PROGRAMA DE DOCTORADO EN SISTEMAS INTELIGENTES

Dra. LOURDES ARAUJO SERNA
Dr. JUAN MARTÍNEZ ROMO

TESIS DOCTORAL

2017



WORD SENSE DISAMBIGUATION IN MULTILINGUAL CONTEXTS

ANDRÉS DUQUE FERNÁNDEZ
M.Sc. in Computer Science and Technology

PROGRAMA DE DOCTORADO EN SISTEMAS INTELIGENTES

Dra. LOURDES ARAUJO SERNA
Dr. JUAN MARTÍNEZ ROMO

A mi madre

AGRADECIMIENTOS

Esta tesis se ha gestado a lo largo de cuatro años que representan una etapa de mi vida. Son muchas las personas que han contribuido a que este trabajo esté ahora terminado, ya sea por sus aportaciones científicas al trabajo en sí, o por sus aportaciones personales. Al fin y al cabo todo suma, y por ello quiero agradecer a todos aquellos a los que considero que les pertenece un trocito de esta tesis.

A mis directores de tesis, Lourdes Araujo y Juan Martínez, por guiarme con tanto acierto en estos años de travesía y compartir conmigo las desesperaciones y los agobios, pero también las alegrías y los reconocimientos al trabajo bien hecho. Miro a mi alrededor y creo que no podría haber tenido mejores directores, indudablemente por sus conocimientos del tema tratado y de la actividad investigadora en general. Pero también por su calidad humana, su capacidad de empatía y trabajo en equipo y su dejar hacer al pupilo que inicia sus andares en la investigación para que desarrolle todas las habilidades asociadas a este trabajo. Sois dos personas muy distintas pero formáis un equipo muy compatible y eficiente, y os estoy muy agradecido por haberme permitido formar parte de él. También quiero extender mi agradecimiento a José Cuesta, por su inestimable ayuda en los aspectos matemáticos de la tesis.

A todos mis compañeros de LSI, los que han estado y los que están. A Ángel Castellanos, Bernardo Cabaleiro, Agustín Delgado y Javier Rodríguez, el núcleo duro siempre necesario para cambiar la realidad a golpe de cafés y cervezas (las últimas siempre fuera del horario laboral, por supuesto). A Rubén Granados, David Hernández, Garazi Olaziregi, Víctor Peinado, Damiano Spina, José Ramón Pérez, Jorge Carrillo, y por extensión a todos los que han pasado por los despachos de la planta baja en algún momento de estos años y con los que he compartido conversaciones, risas y anécdotas. A los integrantes de la "planta noble", en especial a Víctor Fresno, Juan Cigarrán, Álvaro Rodrigo, y otros compañeros de comidas y momentos. A Yolanda Calero por su simpatía y disponibilidad para cualquier cuestión técnica.

A Mark Stevenson, supervisor de mi estancia en la Universidad de Sheffield, por su paciencia, amabilidad y disponibilidad antes, durante y después del tiempo que pasé en Sheffield. Sus conocimientos han sido clave para la realización de este trabajo, y no podría estar más agradecido por su ayuda y comentarios ofrecidos en la revisión del mismo. A todos mis compañeros del *Department of Computer Science*, en especial a Fred Blain, Roland Roller, Jurica Seva, Ahmet Aker, Kashif Shah y Marina Formicheva. Espero que podamos volver a trabajar juntos a lo largo de nuestras respectivas carreras.

A mis amigos, los de toda la vida y los que han ido apareciendo en mi camino, por todas las vivencias que conforman lo que somos, también hay parte de ellos en lo que hago. A los compañeros de la Universidad, por conseguir que años después el espíritu Cebollita siga tan

vivo como el primer día. Sin todos vosotros este viaje no habría sido igual, y yo no sería la misma persona.

Sobre todo, quiero agradecer a los míos. A mi familia. A mi padre Miguel Ángel y a mi hermano Bruno, pero especialmente a mi madre Gislaine, que aunque no podrá verme presentar la tesis físicamente, sé que siempre se ha sentido y se sentirá muy orgullosa de todo lo que hago. En cada logro que consiga en mi vida (y este es uno de los más importantes), ella estará siempre conmigo.

Y a Ana, que ha estado a mi lado casi desde el principio de esta etapa (aunque en el fondo desde mucho antes), y que me ha apoyado con su presencia, su cariño y sus consejos, sobre todo en los momentos más oscuros. Gracias por cada día.

Andrés Duque Fernández

Madrid, Diciembre de 2016

ABSTRACT

Word Sense Disambiguation (WSD) can be defined as the process of identifying the sense adopted by a polysemic word, that is, a word with different possible meanings, in a particular context within a sentence. This process represents a key aspect of any Natural Language Processing task, given the need of determining without ambiguity the correct meaning of all the words within a text, for an automatic system to be able to understand it and work with it.

In this thesis, we present a research focused on Word Sense Disambiguation in scenarios in which it is possible to make use of information written in different languages. Considering those scenarios, we divide the thesis into two lines of study, depending on the specific WSD tasks that are tackled: Cross-Lingual Word Sense Disambiguation, and multilingual Word Sense Disambiguation in the biomedical domain. In the first task, the main aim is to find the most suitable translation for an ambiguous target term written in a source language (typically English) into a target one. The biomedical WSD task is based on finding the most suitable sense of a term that can refer to many different biomedical concepts.

In order to address the proposed tasks, we use a novel technique based on co-occurrence graphs: through that technique, we are able to transform the unstructured information available in different corpora, into a structured base of knowledge that will be subsequently used for performing the disambiguation itself. This knowledge base is a graph in which nodes represent concepts from a given corpus, and the links between those nodes contain information related to the statistical significance of their co-occurrence, that is, of the appearance of both concepts in the same document of the corpus.

Regarding the first task, multilingual information is inherent to the problem itself, since the objective is to find the most suitable translations of words between different languages. For addressing it, our system makes use of the co-occurrence graphs for representing the knowledge in the target language. Then, the contexts of the ambiguous terms, written in the source language and translated through an automatically created bilingual dictionary, are used as source of information for the co-occurrence graph to perform the disambiguation step. In this line of research we also present a study on the possible bilingual dictionaries needed in this kind of tasks.

Considering the biomedical WSD task, in our research multilinguality is used as an additional evidence for testing whether it is possible to improve the performance of monolingual systems addressing the task. For that purpose, we initially adapt our system for tackling the task under a monolingual perspective (in which the co-occurrence graph is built from a corpus written in a single language). After that, we enhance the graph with information from additional languages, in order to study whether this enhancement leads to an improvement of the results obtained by the system. It is a pioneering research in this field, since no similar studies have

been found in the literature that make use of multilingual information for performing WSD in the biomedical domain.

We have explored many different monolingual and multilingual corpora along the development of this thesis, both written with general purposes and related to a specific domain (in particular, the biomedical domain). We have also studied and compared different algorithms that make use of the co-occurrence graph as a structured knowledge base for performing the final disambiguation. The mathematical hypothesis in which the construction of our co-occurrence graph is based, has been compared to similar techniques, offering better results. Similarly, for each of the considered tasks (Cross-Lingual WSD and biomedical WSD), our system has been compared with other state-of-the-art techniques, obtaining very competitive results.

RESUMEN

La desambiguación del sentido de las palabras se define como el proceso de identificación del sentido que adopta una palabra polisémica, es decir, con varios significados posibles, en el contexto concreto de una oración. Debido a la necesidad de definir sin ambigüedad posible el significado de todas las palabras de un texto para que un sistema automático pueda entenderlo y trabajar con él, la desambiguación semántica representa un aspecto crucial y transversal a cualquier tarea dentro del Procesamiento del Lenguaje Natural.

La investigación realizada en esta tesis doctoral se centra en la desambiguación semántica en escenarios en los que existe la posibilidad de utilizar textos escritos en diversos idiomas. Dentro de estos escenarios, dividimos la tesis en dos grandes campos, en función de las tareas específicas de desambiguación a las que nos enfrentamos: desambiguación bilingüe del sentido de las palabras, y desambiguación multilingüe en el dominio biomédico. En la primera tarea, el objetivo es, dada una palabra con múltiples significados, escrita en un idioma inicial (generalmente inglés), encontrar su traducción más adecuada en un idioma final. La tarea de desambiguación en el dominio biomédico se basa en encontrar el sentido correcto de un término médico que puede apuntar a distintos conceptos concretos.

Para hacer frente a las tareas propuestas, se utiliza una técnica novedosa basada en grafos de co-ocurrencia: a través de dicha técnica se transforma la información no estructurada disponible en diversos corpus, en una base de conocimiento estructurada que se puede utilizar después para realizar tareas de desambiguación. La base de conocimiento es un grafo en el que los nodos representan conceptos del corpus, y los enlaces entre ellos contienen información relacionada con la significancia estadística de su co-ocurrencia, es decir, de su aparición conjunta en un mismo documento del corpus.

En la primera tarea, la información multilingüe es inherente al propio planteamiento del problema, ya que se busca obtener las traducciones más adecuadas de palabras entre varios idiomas. En ella, nuestro sistema utiliza los grafos de co-ocurrencia para representar el conocimiento en el idioma objetivo. Los contextos de las palabras ambiguas, escritos en el idioma original y traducidos gracias a un diccionario bilingüe creado automáticamente, se utilizan como fuente de información para que el grafo de co-ocurrencia realice la desambiguación. En esta línea se presenta también un estudio sobre los diccionarios bilingües necesarios en este tipo de tareas.

En lo que se refiere a la desambiguación en el dominio biomédico, el multilingüismo se utiliza como evidencia adicional para comprobar si es posible mejorar la eficacia de sistemas monolingües en la tarea. Para ello, inicialmente se plantea una adaptación de nuestro sistema para hacer frente a la tarea desde una perspectiva monolingüe (en la que el grafo de co-ocurrencia se construye a partir de un corpus escrito en un único idioma). A continuación, se enriquece el grafo con información procedente de idiomas adicionales, para observar si

este enriquecimiento desemboca en una mejora de los resultados obtenidos por el sistema. Se trata de una propuesta pionera en su campo, ya que no se han encontrado otros trabajos que utilicen información multilingüe para la desambiguación en el dominio biomédico.

A lo largo del desarrollo de la tesis, se exploran múltiples corpus monolingües y multilingües, tanto de propósito general como relacionados con un dominio específico (en concreto el dominio biomédico). También se han estudiado y comparado diversos algoritmos que utilizan el grafo de co-ocurrencia como base estructurada de conocimiento para realizar la desambiguación final. La hipótesis matemática en la que se basa la construcción de nuestro grafo de co-ocurrencia ha sido comparada con otras técnicas similares, ofreciendo mejores resultados. Asimismo, para cada una de las tareas consideradas (desambiguación bilingüe y desambiguación en el dominio biomédico), nuestro sistema se ha comparado con otras técnicas del estado del arte, presentando resultados muy competitivos.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Last night I shot an elephant in my pajamas. How he got in my pajamas, I'll never know.

Groucho Marx

Contents

1.1 Scope of the Thesis and Motivation

The definition of ambiguity, as it can be found in the Oxford Dictionary of English, is "A word or phrase susceptible of more than one meaning; an equivocal expression." Hence, ambiguity can be referred to many different aspects of natural language: from isolated words which can present different senses to complete sentences that will have different meanings depending on the receptor of the message. In the example cite by Groucho Marx that can be found in the title page of this chapter, the comedian creates a joke by taking advantage of the ambiguous nature of any language: he plays with the two elements of the sentence that could be "in his pajamas", since the first sentence is structurally ambiguous. In that sentence, it would be technically correct that both "I" and "an elephant" could be "in my pajamas".

If we consider meaning to be composed by regions in a specific space, a denotation would be the correspondence between an expression and a region of that "meaning space". Hence, an expression is considered to be ambiguous if it has two or more distinct denotations (Wasow et al., 2005). These denotations can be completely disjoint, making the ambiguity more obvious, but they can also share characteristics, which makes the task of determining the correct interpretation of the expression much more difficult.

As we have introduced before, natural language presents many different levels of ambiguity. These types of ambiguity can be classified into the following categories:

- Lexical ambiguity: This type of ambiguity refers to the possible meanings or senses of a simple word. It represents the lowest level of ambiguity, and can be subdivided into two types: First, words that are ambiguous with respect to its syntactic class, e.g., the word *"silver"* which can be used as a verb, noun or adjective. The second type of lexical ambiguity would be lexical-semantic ambiguity, in which a single word is associated with multiple senses (bank, pen, bat, etc.).

  (1.1) *I saw a bat.*

  (1.2) *She put the money in the bank.*

  Example 1.1 is a case of lexical-semantic ambiguity, in which the word *"bat"* could refer to a flying mammal or to a wooden club. The lack of more context in the sentence makes it impossible to determine the correct sense of the ambiguous word. Example 1.2 is another case of lexical-semantic ambiguity. Here, the word *"bank"* could mean "side of a river", or "financial institution". In this case, context gives us enough information to infer that the most suitable sense of the word *"bank"* would be "financial institution". However, even if we considered the alternative definition of *"bank"*, the sentence would still be correct, since she could have hidden the money in the bank of a river.

- Structural ambiguity: Also known as syntactic ambiguity or grammatical ambiguity, it arises from the arrangement of words and structures in sentences, or from the classification of words. It involves the presence of two or more different meanings for a complete sentence or sequence of words.

  (1.3) *Every man loves a woman.*

  (1.4) *The girl saw the boy with the telescope.*

  In Example 1.3, the syntactic ambiguity is called scope ambiguity, since it involves operators and quantifiers. In this case, the receptor of the message could interpret that there is a different woman for every man to love her, or that every man loves the same woman. Example 1.4 illustrates attachment ambiguity, since there exists a constituent which fits more than one position in a parse tree. In this case, depending on where the receptor attaches the preposition "with" (either to "the girl" or to "the boy"), the sentence will mean that the girl used a telescope to see the boy, or that the girl observed a boy who was using a telescope.

- Semantic ambiguity: It is the ambiguity that remains even after the syntax and the meanings of individuals words have been resolved, due to the fact that the meaning of the words themselves can be misinterpreted.

  (1.5) *John and Mary are married.*

  In Example 1.5, the sentence presents a semantic ambiguity, since it could be interpreted in two different ways: either John and Mary are married to each other, or they are married separately, each of them to a different partner.

- Anaphoric ambiguity: It is a type of ambiguity that involves deeper knowledge of the context, since the interpretation must be carried out using this knowledge. It arises when some entities previously introduced in the discourse are referred to in subsequent situations.

  (1.6) *Bill told Amy that he had decided to spend a year in Italy to study art.*
  That would be his life's work.
  After he had done that, he would come back and marry her.
  That was the upshot of his thinking the previous night.
  That started a four-hour fight.

  In Example 1.6, the word "that" introduces ambiguity in all the sentences that follow the first one, and the resolution of this ambiguity requires a deep knowledge of the context of each sentence. In the second sentence, "that" can be replaced by "art". In the third sentence, "that" refers to "spending a year in Italy". The fourth "that" can be interpreted as "the whole decision". Finally, the word "that" in the last sentence, refers to "telling Amy everything he had thought".

In this thesis, we will focus on the first type of ambiguity, lexical and lexical-semantic ambiguity, that is, words that can present more than one meaning. We will work with both words whose meaning changes depending on the part of speech, and words with different senses even falling into the same part of speech.

Word Sense Disambiguation (WSD) is the task that attempts to automatically solve this lexical and lexical-semantic ambiguities by determining the most suitable sense for an ambiguous target word given its context. Although it is not usally considered as an end product, but rather a component within another system (Preiss, 2006), it has become one of the major issues of Natural Language Processing (NLP), due to its importance for many other NLP processes. Some of those processes that can benefit from WSD are the following (Ide and Veronis, 1998):

- Machine translation, related to Cross-Lingual Word Sense Disambiguation: the task of determining the most suitable translation of an ambiguous target word, from an initial language to a final one.
- Information retrieval and hypertext navigation, for eliminating non relevant results.
- Content and thematic analysis, such as text classification or text summarization.
- Grammatical analysis, in the process of improving part of speech tagging or preposition phrase attachment.
- Speech processing, for the correct phonetization of words, word segmentation, homophone discrimination.
- Text processing: diacritics insertion, case changes (e.g., discriminating between proper and common nouns).

WSD has been frequently treated as a supervised learning problem (Màrquez et al., 2006; Mihalcea, 2006), based on techniques that depend on scarce and expensive resources such as semantically tagged corpora or lexical databases like WordNet (Fellbaum, 1998), which will be detailed in subsequent chapters of this thesis. Unsupervised techniques, on the contrary, do not require semantically annotated resources, and are commonly known as Word Sense Induction (WSI) techniques. Their objective is to induce the different senses of a specific word in a given text by selecting groups of words related with a particular sense of the word. Those techniques are often divided into vector-based and graph-based techniques. Graph-based techniques (Agirre et al., 2006b; Klapaftis and Manandhar, 2008; Veronis, 2004) are directly related with the work that will be developed in this thesis. In Chapter 2, we will analyse in more detail different systems that can be found in the literature, addressing both WSD and WSI under various perspectives. As we will see in that chapter, most of the graph-based approaches are based on the concept of co-occurrence for creating links between nodes in the graph: they usually represent each word co-occurring with the target word to be disambiguated, inside a predefined window, as a vertex in the graph. For the sake of understanding, we need to properly define the concept of co-occurrence: It

is a linguistic term that can either mean concurrence / coincidence or, in a more specific sense, the occurrence of two terms from a text corpus alongside each other in a certain order. Co-occurrence in this linguistic sense can be interpreted as an indicator of semantic proximity or an idiomatic expression. In contrast to collocation, co-occurrence assumes interdependency of the two terms. However, for the purposes of our work, we will not take the order of occurrence of the two terms in a co-occurrence into account, but only the fact that they are happening at the same time in the same context, which will be defined later on.

The initial hypothesis for our work is devoted to the development of unsupervised graph-based systems for performing WSD:

WORK HYPOTHESIS 1: ***Unsupervised methods based on co-occurrence graphs reduce the amount of annotated external resources to a minimum, and are able to achieve competitive performances in diverse Word Sense Disambiguation tasks.***

Following this hypothesis, an unsupervised graph-based technique for performing Word Sense Disambiguation tasks will be used and deeply analysed throughout this thesis. Its results in different tasks will be compared to state-of-the-art techniques in order to prove its validity.

The technique is based on the work by Martinez-Romo et al. (2011), and its mathematical foundations will be detailed in Chapter 3. Its main objective is the transformation of unstructured information that can be found in resources such as corpora into a structured representation of knowledge. This structured representation is a co-occurrence graph, that is, a graph that contains information about the co-occurrences between the concepts represented in its nodes. The co-occurrences are extracted from text provided in those unstructured sources of information (textual corpora). As we stated before, most of the methods based on co-occurrences consider a pre-defined window for analysing those co-occurrences, that is, only those words inside the window are considered for extracting co-occurrence information. The main difference of the technique presented here is that it considers a complete document to be that co-occurrence window. For the purposes of the technique, we consider a document to be a coherent piece of information, so we can assume that elements (words or annotated concepts) in a document tend to (statistically) adopt a related sense. Hence, we consider that two elements frequently appearing in the same document, that is, frequently co-occurring, share some kind of semantic information that points to a related sense, and must be connected by an edge in a graph. The validity of this assumption is another secondary hypothesis of this thesis, related to the hypothesis WH1.

The construction of the graph is then based on a mathematical hypothesis that considers the statistical significance of the co-occurrence of two elements inside the same document. This significance is extracted by comparing the real co-occurrences of the elements against a null model that represents what could be considered co-occurrence by pure chance. A more detailed description of the mathematical background that supports the subsequent experiments carried out along the development of the thesis can be found in Chapter 3 of

this document. Knowledge represented by the co-occurrence graph is used for performing Word Sense Disambiguation. This can be done through many algorithms that extract that information and use it to determine the sense associated with a word within a specific context. Those algorithms can be oriented to find subgraphs (communities) inside the co-occurrence graph, each of them representing a different sense. Also, other algorithms that highlight the importance of a node (word) in the graph, depending on the context that surrounds it, will also be tested in order to determine their suitability for the tackled problems.

Apart from this initial hypothesis, a second foundation of our work is the idea of multilinguality, and the improvements that can be achieved by integrating multilingual resources in Word Sense Disambiguation tasks. In our work, parallel and comparable corpora will be considered the source of knowledge to perform the disambiguation. These corpora are good resources for NLP in general (Resnik, 2004), since parallel translations share hidden meaning that can be useful for extracting knowledge about a language, from another language richer in resources.

WORK HYPOTHESIS 2: ***The use of multilingual resources such as parallel and comparable corpora provides a qualitative enhancement of information to our Word Sense Disambiguation techniques, which can be reflected in an improvement of the overall results.***

For testing and proving this hypothesis, we will address the task of using multilinguality for WSD under two different perspectives:

- **Cross-Lingual Word Sense Disambiguation:** First, we will consider an eminently multilingual task, such as Cross-Lingual Word Sense Disambiguation. CLWSD, which is highly related to machine translation processes, aims to determine the most suitable translation for a given word from a source language to a target one. Some of the difficulties of WSD, such as the scarcity of sense inventories and sense tagged corpora, are particularly important in this task. CLWSD systems usually take advantage of the shared meaning between parallel texts for dealing with those issues. Parallel corpora are hence considered as the main source of knowledge for performing disambiguation in this field. The disambiguation requires another resource that proposes a set of possible translations for the system to choose the most suitable ones. This resource will be a bilingual dictionary between those languages that take part in the disambiguation. This dictionary offers the potential translations of the target word, as well as translations for the words surrounding the target word (context). For each test instance containing a target word, any system performing CLWSD will have to choose its most suitable translation, among those offered by the dictionary. The choice of an adequate bilingual dictionary is one of the most important decisions to be made along the disambiguation process, in order to ensure the good performance of a CLWSD system. An analysis of the different types of bilingual dictionaries that can be used for CLWSD will be also carried out as a secondary problem inside the CLWSD task.

- **Word Sense Disambiguation in the Biomedical Domain:** We have selected a specific domain for testing our technique, and the second hypothesis presented here. The main goal is to explore whether the use of multilinguality can offer significant improvements compared to monolingual systems addressing biomedical WSD. Hence, we will study the results that can be achieving by enhancing the knowledge base used in the disambiguation process (the co-occurrence graph) through multilingual techniques. The vast amount of unstructured textual information available in the biomedical sciences has created the need for automatic systems to access, retrieve and process these documents (Savova et al., 2008). However, this is made more difficult by the range of lexical ambiguities that are found in these documents, including different meanings of general terms or the different extended forms of acronyms and abbreviations. For example, the word "surgery" may refer to the branch of medicine that applies operative procedures to treat diseases, or to one of those operative procedures. Also, the acronym "BSA" could refer to multiple expansions such as "Bovine Serum Albuminum" and "Body Surface Area". It is difficult to find works in the literature that apply multilinguality to the WSD task in the biomedical domain, probably due to the lack of bilingual corpora providing enough useful information for disambiguation, that is, a wide enough collection of documents containing ambiguous terms, and with a balanced number of occurrences for each possible sense of such terms. Yet, WSD is of paramount importance for many document processing tasks in this field, such as summarisation, text classification or information extraction. New possibilities of improvement related to WSD are thus highly relevant in the NLP field.

The problem of CLWSD will allow us to start our study on the application of multilingual information in WSD tasks, since it is a problem that can only be addressed from a multilingual perspective, and the use of multilingual resources is mandatory for successfully approaching the task. The motivation for studying the second problem, WSD in the biomedical domain, comes from our will of analysing whether multilingual information is also useful in a task which has been historically addressed from a monolingual perspective. We will try to prove our second work hypothesis by determining the usefulness of those multilingual resources in this task, and the circumstances under which significant improvements can be achieved.

1.2 Research Objetive and Research Questions

As we have briefly introduced in the past paragraphs, the main goal of this thesis is to tackle the problem of Word Sense Disambiguation under a multilingual point of view, both in general and specific domains. We can formalize this objective in the following statement:

RESEARCH OBJECTIVE: ***Study the problem of Word Sense Disambiguation in the scope of Natural Language Processing, and the importance of solving this problem also in specific domains such as biomedicine. Analyse the usefulness of multilinguality to improve systems performing WSD and develop an unsupervised graph-based system able to overcome state-of-the-art techniques in different WSD tasks.***

This main goal can be divided into smaller objectives, each of which will be represented as a research question. These questions will become the central axis of the thesis, and hence all the decisions taken along the development of the thesis will aim to answer them.

- **General Word Sense Disambiguation**
  - **Research Question 1:** *Considering the idea of coherence inside a document, is the co-occurrence graph a valid structured representation of the information inside a corpus, for addressing Word Sense Disambiguation tasks?*
  - **Research Question 2:** *Once that the information has been formally represented in a co-occurrence graph, which disambiguation algorithm makes better use of this structured source of knowledge to perform disambiguation?*
  - **Research Question 3:** *How close can get the results obtained by an unsupervised Word Sense Disambiguation algorithm to those achieved by a supervised one?*
  - **Research Question 4:** *How can we combine multilingual information available in parallel and comparable corpora with our proposed technique for performing Cross-Lingual Word Sense Disambiguation?*
  - **Research Question 5:** *Is our unsupervised graph-based technique able to overcome other state-of-the-art approaches in Cross-Lingual Word Sense Disambiguation tasks?*
  - **Research Question 6:** *What is the impact of selecting different bilingual dictionaries in a system performing Cross-Lingual Word Sense Disambiguation?*
- **Word Sense Disambiguation in the Biomedical Domain**
  - **Research Question 7:** *Can we successfully apply our original disambiguation technique to Word Sense Disambiguation in the biomedical domain? Which are the adjustments that should be applied to our algorithm for tackling the WSD task in this domain?*
  - **Research Question 8:** *Can unsupervised Word Sense Disambiguation be improved by multilingual information in the biomedical domain? Under which circumstances?*
  - **Research Question 9:** *Is multilingual data usually available in the biomedical domain? If not, which is the best way to automatically supply multilingual information to a system performing WSD in this domain?*

We have divided our research questions into two separated, although strongly connected, research lines. The first one refers to general Word Sense Disambiguation tasks, and specifically Cross-Lingual Word Sense Disambiguation, which aims to find the most suitable translation of an ambiguous word from a source language to a target one. The second research line will be Word Sense Disambiguation in the biomedical domain, in which the main goal is to find the most suitable sense of an ambiguous biomedical concept.

Despite this initial division, we can perform a more thorough classification of the research questions described above: RQ1, RQ2 and RQ3 refer to the developed technique, its validity and the possibility of outperforming state-of-the-art systems in WSD tasks. Information for answering these first two research questions will be gathered from the study of general WSD problems, and from the study of biomedical WSD. RQ4, RQ5 and RQ6 are more related to the general WSD task that we have selected for the first part of the thesis: Cross-Lingual Word Sense Disambiguation. The importance of multilingual resources will be analysed for answering these three research questions, as well as specific characteristics of the addressed task such as the importance of the bilingual dictionaries. Finally RQ7, RQ8 and RQ9, fall in the category of biomedical WSD: the application of our techniques to the biomedical domain, and the study of multilingual resources and the potential improvements that can be achieved by using them will be crucial for answering these three last research questions.

1.3 Methodology

In this section we detail the nature of the different steps followed along the development of this thesis, being the final objective of this methodology to answer all the research questions presented in Section 1.2. Each of the following steps has been followed to some extent, for every task that has been faced during the thesis.

1.3.1 Analysis of Previous Work

A thorough analysis of the available works in the literature has been conducted in order to gather all the possible information about the state of the art of every research challenge that has been addressed. We have explored both supervised and unsupervised techniques. This exhaustive analysis has a double purpose: First of all, we need to make sure that our intended work has not been previously explored by other researchers; in case similar approaches are found in the literature, it is important for us to identify which parts of those previous works can be potentially useful for improving our system. The second objective of the literature analysis is to know all the latest state-of-the-art systems tackling the same problem that we address, in order to compare them with our own techniques. Through this analysis we obtain

values of accuracy and performance of those state-of-the-art systems, defining this way the minimum performance that our system should achieve in order to consider it suitable for addressing the proposed problems.

All the information gathered in this stage, for every task considered in this work, has been summarized in Chapter 2.

1.3.2 Sources of Information

Depending on the task that we intend to address (Cross-Lingual Word Sense Disambiguation, monolingual or multilingual Word Sense Disambiguation in the biomedical domain, etc.), the available sources of information will change. Considering that we are facing problems in the field of Natural Language Processing, these sources of information will usually be corpora, i.e., collections of unstructured text documents from where we expect to extract the knowledge contained in these documents. Once formally represented in graphs, this knowledge will allow us to perform WSD in an automatic way.

Although there exist different corpora that could be useful for our purposes, it is important that we carry out an exhaustive evaluation of the possibilities that each of the corpus offer, in order to be sure that the selected ones fulfill our work hypothesis (i.e., they present semantic coherence at document level). As we stated before, the fundamental nature of the selected corpora will change depending on our needs, and hence we will work with monolingual and multilingual general purpose corpora such as Europarl (Koehn, 2005), or monolingual and multilingual biomedical corpora such as NLM-WSD (Weeber et al., 2001) or EMEA (Tiedemann, 2009).

The generation of our own corpora is also a key step in our methodology. At some stages of our research, we have dealt with the lack of useful corpora for our purposes. This is the reason why we have decided to eventually create two corpora, both inside the biomedical domain: First, a monolingual (English) corpus with documents downloaded from PubMed and containing, each of them, at least one of the ambiguous words of the test dataset NLM-WSD. The second corpus is a bilingual (English-Spanish) collection of biomedical documents extracted from Ibero-american journals, referred to rare diseases, and also containing words from the NLM-WSD test dataset.

1.3.3 Representation

The next step in our methodology is the adaptation of our algorithm for building the co-occurrence graph to the specific problem we are facing. Knowledge representation of the

information contained in corpora using co-occurrence graphs is one the most important contributions of this thesis, and is a crucial process for a successful application of disambiguation algorithms. As we will explain in further chapters of this document, the definition of the specific disambiguation task that we tackle will provide us with enough information to determine which elements will eventually become nodes of our graphs. That is, depending on the nature of the task, we will choose between direct textual information (words such as nouns, verbs or adjectives), annotated semi-structured information such as identifiers, or both types of elements, as nodes of the co-occurrence graphs.

The variation of elements that populate the co-occurrence graphs has not only been explored when facing different WSD problems, but we have also tested different representations for the same task. For example, in multilingual WSD in the biomedical domain, representations involving only textual information and combining it with concept identifiers have been analysed. In other tasks, despite of using only textual information, we have also varied the part-of-speech elements used for building the co-occurrence graphs, and analysed the different results achieved by the technique in each case.

1.3.4 Algorithm Design

As we stated before, a correct structured representation of the available knowledge for performing the disambiguation is crucial in our methodology. However, the way we use these data structures (in this case, co-occurrence graphs) will determine the success or failure of our system. For each problem we need to define algorithms which, given new input instances, perform the disambiguation of target words using their context as source of information. These algorithms will have to play with the relations between elements of the co-occurrence graph, and the weights of those relations, for providing a ranking containing all the possible senses of the ambiguous target word, ordered by their likelihood of being the most suitable sense in that instance.

As we will explain in further chapters of this document, there exist many different graph-based algorithms that can be used in this step. In particular, we are specially interested in the comparison between algorithms that rely on the structural nature of the graph and algorithms that are more focused on the direct relations between words and the weights of those relations. In the first category, we can consider algorithms such as PageRank, or community extraction algorithms (Walktrap, Chinese Whispers). The second type of algorithms are more related to finding the shortest paths between nodes in the graph (Dijkstra's algorithm and its variations).

1.3.5 Evaluation

The evaluation of the proposed approaches is highly related to the selection of the specific WSD tasks we want to tackle. That is, when we analyse which problems can be faced using our techniques we also need to determine how the evaluation of the performance of those techniques should be done. In this process, also close to the analysis of available corpora explained in Subsection 1.3.2, we need to explore the availability of gold standards against which we will match the results offered by our system, as well as possible baselines for each of the faced problems.

Results obtained by our system in the considered evaluation frameworks should be discussed and thoroughly analysed, both in an isolated way, that is, studying the conclusions which can be derived regarding our own system, but also by comparing them with other state-of-the-art systems tackling the same task. We will usually select knowledge-based unsupervised systems as we introduced in Subsection 1.3.1 for comparing results in a fair manner, although it is important that we also take into account supervised state-of-the-art systems, in order to get insights about the answer to our research question RQ3.

1.3.6 Summarization and Dissemination

The last step of our methodology has to do with the publication of achieved results and the conclusions obtained through their analysis. Apart from this dissertation, which comprises all the aspects of the developed work, we have published the most relevant parts of the research in important conferences and journals. These publications are summarized in Section 8.4 of Chapter 8.

1.4 Structure of the Thesis

The rest of this document is structured as follows:

- Chapter 1, Introduction: We present the motivation for the work on Word Sense Disambiguation in multilingual contexts, detailing the main problem we want to tackle, the work hypothesis, research objective and research questions, as well as the methodology followed along the development of the thesis.

- Chapter 2, Related Work: We provide a study on the most relevant works in the field. The different paradigms for addressing WSD are presented, and the latest systems addressing both Cross-Lingual WSD and biomedical WSD are analysed.
- Chapter 3, Theoretical Background: We describe the foundations of the proposed technique, as well as an empirical comparison with other similar co-occurrence techniques.
- Chapter 4, Cross-Lingual Word Sense Disambiguation: We present our study on Cross-Lingual Word Sense Disambiguation, specifically describing the adaptation of our technique for addressing that task. Results obtained by our system, and a comparison with other state-of-the-art systems, are detailed.
- Chapter 5, Dictionaries for Cross-Lingual Word Sense Disambiguation: An exhaustive study on the impact of choosing different bilingual dictionaries for performing Cross-Lingual Word Sense Disambiguation tasks is carried out in this chapter. Four different dictionaries are presented, and results obtained by an ideal system and by our graph-based technique are analysed.
- Chapter 6, Word Sense Disambiguation in the Biomedical Domain: We adapt the proposed technique for its application to Word Sense Disambiguation tasks in the biomedical domain, under a monolingual perspective. We detail the adjustments that need to be done to the proposed technique, as well as the achieved results in two widely known test datasets.
- Chapter 7, Multilinguality for Biomedical Word Sense Disambiguation: We introduce the use of multilinguality for performing biomedical Word Sense Disambiguation. New multilingual resources are presented, based on manual and automatic translations. Results obtained by our technique, enhanced with multilingual information, are finally described.
- Chapter 8, Conclusions and Future Work: The main contributions and conclusions of the work are summarized, and future lines of research are also analysed. Finally, we present the publications related to the development of this thesis.

2

RELATED WORK

To acquire knowledge, one must study; but to acquire wisdom, one must observe.

Marilyn vos Savant

Contents

This chapter is devoted to the analysis and description of previous works related to the main lines of research of this thesis. A brief history on Word Sense Disambiguation is presented in Section 2.1, as well as a survey on the most important techniques and systems developed in the last years for addressing WSD, regarding the different categories into which an approach can be classified. The use of multilinguality for enhancing the available knowledge when facing these tasks, and WSD in particular, is presented in Section 2.2. Finally, we illustrate the latest developments on Word Sense Disambiguation in the biomedical domain in Section 2.3.

2.1 Word Sense Disambiguation

2.1.1 The WSD Task

Word Sense Disambiguation is the task which aims to computationally determine the correct sense of an ambiguous word inside a context. Given a text T represented by a sequence of words $(w_1, w_2, ..., w_n)$, WSD tries to find the correct sense to all or some of those words, that is, a mapping A from words to senses, where $A(w_i) \subseteq Senses_D(w_i)$, being $Senses_D(w_i)$ the set of possible senses for word w_i that can be found in a dictionary D. The WSD task is a classification task in which the possible senses for a word would represent the classes. The difficulty of the task comes from the fact that the number of classes is variable, that is, for each word there exist a different number of possible senses. Hence, WSD can be seen as a set of n classification tasks, being n the size of the lexicon (number of potential words to be disambiguated) (Navigli, 2009).

Being always a subprocess for more complex NLP problems, WSD has been tackled from the late 40s and early 50s. The importance of context and a study on the minimal size of the context window that should be analysed for unequivocally determining the correct meaning of an ambiguous word is initially presented in (Weaver, 1955). Artificial Intelligence (AI) techniques such as semantic network and their use for language understanding were deeply used in the 60s and 70s for addressing the WSD problem (Masterman, 1961; Wilks, 1968), and particularly the use of WSD for Machine Translation (MT). The creation and releasing of large-scale resources such as dictionaries, corpora or lexicons in the 80s and 90s had the effect of relieving the knowledge acquisition bottleneck (Gale et al., 1992a), that is, the problem of correctly gathering the huge amount of necessary knowledge for performing WSD (dictionaries, semantic network, annotated corpora, etc.) and representing this knowledge in a formal manner which could be used by automatic systems.

The approaches for performing WSD can be categorized into three different types: supervised, unsupervised and knowledge-based. However, there are cases in which a system can be classified as belonging to different categories, depending on the algorithms used and the resources involved in the disambiguation process. In the next subsections, we show the most important techniques and systems for each of the categories mentioned above.

2.1.2 Supervised Methods

In a similar way to many other techniques and algorithms related to machine learning and data mining, supervised methods represent a very important line of research when it comes to WSD, as they are usually able to achieve the best overall results. However, these systems usually rely on annotated data for performing a training phase in which the parameters of a model are learned. In the case of WSD, supervised techniques need semantically annotated texts in which the correct sense of every target ambiguous word is already selected. Obtaining this kind of resources is an expensive process, which is the reason why this kind of corpora are scarce and difficult to obtain. Some examples of sense annotated corpora are *SemCor* (Miller et al., 1993) or DSO (Defense Science Organisation) (Ng and Lee, 1996).

Supervised learning for Word Sense Disambiguation can be seen as a two-step process: First, a set of features for representing the context of an ambiguous target word needs to be defined. These features will become the input for a system when a new instance needs to be disambiguated. The second step is the learning algorithm itself, which will be trained with the sense annotated corpus in order to learn the parameters of the model (Martínez et al., 2002). The context features are usually divided into local features, which refer to the most immediate context of an ambiguous term (*n-grams* containing lemmas and Part-Of-Speech tags from words inside a small window surrounding the target word), and topical features, considered as a bag of words from a more general context, such as the sentence containing the target word, and even previous and subsequent sentences (Agirre and Edmonds, 2007a).

Considering the second step of the supervised WSD process, many different machine learning algorithms have been used for performing the final disambiguation. Techniques such as Decision Trees (Quinlan, 1986), Decision Lists (Rivest, 1987), probabilistic classifiers, exemplar-based learning or Support Vector Machines (Boser et al., 1992) have been widely used in the literature, as well as their combination using boosting methods. Exhaustive comparisons of supervised methods can be found in different surveys (Agirre and Edmonds, 2007a; Lee and Ng, 2002).

In the past few years, the field of representing the context as features for performing supervised WSD has witnessed various breakthroughs with the incorporation of word embeddings. This technique was initially presented by Bengio et al. (2003) and subsequently improved by Collobert and Weston (2008) and specially by Mikolov et al. (2013) with the creation of the model *Word2Vec*. Word embeddings can be defined as a representation

paradigm for transforming the semantic space of words into a real-valued continuous vector space (Iacobacci et al., 2016). The technique for creating word embeddings, based on deep neural networks, aims to create a set of vectors in a low-dimensional space in relation with the size of the vocabulary of the corpus used for training the model. These vectors can be then used as features for many different NLP processes, for example WSD. The work by Iacobacci et al. (2016) proves the usefulness of word embeddings for performing supervised WSD by incorporating sets of features based on this technique to a widely known supervised system: It makes sense (IMS) (Zhong and Ng, 2010). This system uses Support Vector Machines as base algorithm for performing the disambiguation, although it is a widely flexible framework that allows the modification of its main modules: pre-processing, feature extraction and classification.

The creation of word embeddings can be seen as an unsupervised process in relation to the disambiguation itself, since sense annotated corpora are not needed for generating the vectors, which are automatically created from raw texts. That is the reason why some authors consider the combination of supervised WSD systems such as IMS with word embeddings to be a semi-supervised approach (Turian et al., 2010). For example, in Taghipour and Ng (2015), word embeddings initially obtained in the work by Collobert and Weston (2008) are adapted to the specific domain through a neural network that incorporates discriminative information about the task and used for performing WSD also through the IMS system. The main idea beyond the generation of word embeddings can be extended (Rothe and Schütze, 2015) for incorporating information from lexical resources such as WordNet (Fellbaum, 1998) or knowledge bases like Wikipedia or Freebase (Bollacker et al., 2008). Finally, other approaches based on neural networks have been also applied to the problem of WSD. In Yuan et al. (2016), a language model is created through a Long-Short Term Memory based neural network (Hochreiter and Schmidhuber, 1997) for predicting a word from its surrounding context. This language model is then used for extracting the context vector of a test instance. A vector for each possible sense of the target word has been previously created using the context vectors of example sentences of that sense. The most appropriate sense for the target word in the test instance will be the one whose vector presents maximal cosine similarity with the context vector of the test instance.

2.1.3 Unsupervised Methods

Unsupervised Word Sense Disambiguation is usually known as Word Sense Induction (WSI), and instead of selecting the most appropriate sense of an ambiguous word among a closed list of discrete sense candidates, it considers a more complex nature and distribution of the senses of a word. Hence, the main aim beyond algorithms performing unsupervised WSD is to induce the possible senses of every target ambiguous word directly from a corpus, typically by grouping (clustering) similar examples which will eventually represent the different senses of ambiguous words. The evaluation of these clusters is a key part of unsupervised

WSD, and can be done under three different approaches: by manually determining the appropriateness of the clusters for representing each of the possible senses of the target word, by developing a method for automatically mapping the clusters to senses of the word provided by a lexicon, or by embedding the WSI module into a more complex NLP system which performs tasks such as information extraction or text summarization, and evaluating the whole system (Agirre and Soroa, 2007).

The sense induction process usually falls into three different categories: context clustering, word clustering and co-occurrence graphs.

- **Context clustering**: The objective of this approach is to extract clusters based on the context of each target word. For this purpose, for each occurrence of a target word, its context is transformed into a vector. These context vectors are then clustered into different groups, each of them will represent a sense of the target word, through algorithms such as the context-group discrimination algorithm (Schütze, 1998). As we stated in Subsection 2.1.2, the development of word embeddings for representing words as vectors has been widely studied these past few years. However, the original concept of word embeddings represents each word as a unique vector, without taking into account the different possible senses of words. The latest works based on context clustering design context-dependent vector representations of ambiguous words, in which different senses of the word are represented by different vectors (Neelakantan et al., 2015; Reisinger and Mooney, 2010). The meaning of a word is represented by a set of different vectors, so the most appropriate sense of a target word given a context can be selected by choosing the sense vector of the set which minimizes the distance to the context vector of the test instance.

- **Word clustering**: Unlike context clustering, this approach aims at clustering semantically similar words into groups that will represent different senses. It relies in different similarity measures for determining the distance between two different words, and by considering that semantic distance, groups the words using different algorithms. In one of the first works on word clustering by Lin (1998), semantic similarity between pairs of words is computed using the information contained in their shared features, which are extracted from the syntactic dependencies between those features, so the similarity between two words is directly proportional to the number of features they share. Once the similarities are calculated, the different clusters for a word w_i are developed through a similarity tree, in which a word w_j will become a direct child of w_i if their similarity value is higher than the similarity between w_j and another descendant of w_i already in the tree. Each direct child of w_i and the branch of words that grows below it will represent a different sense of that w_i. The similarity measure was also used in the development of the *Clustering by Committee* algorithm (Pantel and Lin, 2002).

- **Co-occurrence graphs**: A new set of techniques for performing sense discrimination and unsupervised WSD has been recently developed. The main data structure beyond

the technique is called co-occurrence graph: A co-occurrence graph is a graph $G = (V, E)$ in which vectors or nodes $v \in V$ correspond to words inside a text and edges $e \in E$ between nodes indicate the simultaneous occurrence (or co-occurrence) of two words in a given context of arbitray size (sentence, paragraph, document,...). Many works can be found in the literature of the past few years developing graph-based algorithms for extracting sense inventories of ambiguous words in a Word Sense Disambiguation pipeline. There exist important differences in the criteria taken by different works for determining the words populating the graph of co-occurrences, as well as for weighting the importance of a co-occurrence by assigning a score to the edge that represents it. In the work by Widdows and Dorow (2002) only nouns that co-occur separated by the words *and* or *or* are taken into account, and the complete number of co-occurrences is considered as a weight for the edge. In that work, the final clustering of nodes is performed by adding the "most similar node" to each cluster, which is the node that, being a neighbour of the cluster (that is, connected to any of the nodes already in the cluster), presents the highest number of connections with other neighbours of the cluster. In a similar way, Hope and Keller (2013) build the co-occurrence graph for each target word by extracting co-occurrences of nouns from coordination patterns in text, although the weights for the links between words are calculated using the Log-Likelihood Ratio. Other graph clustering algorithms such as Chinese Whispers (Biemann, 2006) or Walktrap (Pons and Latapy, 2005) will be explored and detailed in further chapters of this document.

One of the most important breakthroughs in the study of co-occurrence graphs is the *HyperLex* system (Veronis, 2004). In this work, a subcorpora is created for each of the target words, and co-occurrence information is extracted from this subcorpora for each pair of words A and B: f_A and f_B are the occurrence frequencies for each word, respectively, while $f_{A,B}$ is the number of co-occurrences of those two words. Then, the weight for a link between A and B in the graph is calculated as $w_{A,B} = 1 - \max[p(A|B), p(B|A)]$. The probability $p(A|B)$, that is, the probability of finding the word A in a given context knowing that it contains B, is $p(A|B) = f_{A,B}/f_B$. Similarily, $p(B|A) = f_{A,B}/f_A$. Once that the graph is built, the technique explores the idea of "High Density Components" or highly interconnected nodes of the graph, for extracting *hubs* related to the target word. These hubs are the most important nodes in the graph, regarding their degree, and represent the main senses of the target word. An iterative process is carried out for detecting these hubs and deleting them, together with their neighbours, until a criterion related to the degree of the remaining nodes is reached. Finally, a Minimum Spanning Tree (MST) is calculated for the graph, forcing the target word to be the root of the tree, and the hubs detected in the previous step to be directly linked to this root, with weights equal to zero. Figure 2.1 shows an example of the appearance of the co-occurrence graphs created in the *HyperLex* system, and the conversion to a Minimum Spanning Tree in order to perform the disambiguation.

The disambiguation step in the *HyperLex* system is based on the MST built in the last step. Given a context $W = w_1, w_2, ..., w_n$, each context word w_j (different from

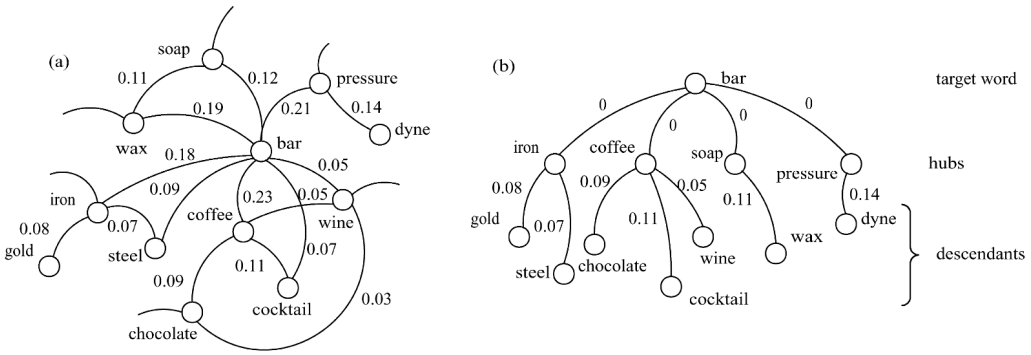


Figure 2.1: (a) Co-occurrence graph in the *HyperLex* system. (b) Minimum Spanning Tree (MST) for the target word *bar* (Navigli, 2009).

the target word) will present an associated score. This score is a vector with as many components as hubs (senses) for the target word, and each of the component of the vector, s_k is weighted depending on the existence of the context word w_j in any of the descendants of the hub which represents, and the distance in the MST between w_j and the hub h_k. This way, each context word will be weighted by its relation to the different senses of the word. Finally, score vectors for the context words are summed up, and the hub with highest score is selected as the most appropriate sense for that target word in the given context. As we can observe in Figure 2.1, depending on the appearance of words classified as "descendants" in the context of a particular test instance, the algorithm will select one of the four different senses selected to be hubs of the target word *bar*. The performance of this technique for Word Sense Disambiguation tasks has been deeply analysed in different works (Agirre et al., 2006a), and compared to another important graph-based algorithm such as PageRank. The initial PageRank algorithm, developed by Brin and Page (1998), was applied for finding the most important web pages related to a specific query, and is known to be the basis of the Google search engine. However, the mathematical background behind it makes it appliable to many other processes which rely on graphs for representing information. In this case, it is possible to detect the most important nodes in a co-occurrence graph and select them as hubs, or senses of the target word (Agirre et al., 2006b). The PageRank algorithm and its variants for Word Sense Disambiguation will be detailed in subsequent chapters of this document.

2.1.4 Knowledge-Based Methods

A third main category of systems for performing Word Sense Disambiguation covers many different methods which do not directly rely on sense annotated corpora for training models, but make use of external resources containing information about the possible senses of an ambiguous target word. Although the performance achieved by this kind of methods is usually below supervised methods, the coverage they obtain is higher. Hence, they are specially useful for *all-words* WSD tasks (Mihalcea, 2006), that is, tasks in which the desired output is the correct disambiguation of all the words in a test instance, in opposition to *lexical sample* WSD tasks, in which only a specific target word has to be disambiguated (normally words that can be found in semantically tagged corpora).

As we have stated in the previous sections, every system performing WSD relies on resources. However, in the case of supervised WSD and unsupervised WSD (or WSI), these resources are unstructured, usually corpora of documents containing text, which can be sense-tagged (in the case of supervised WSD) or raw (for performing WSI). When it comes to knowledge-based WSD, on the other hand, the external resources on which systems rely are structured sources of information: thesaurus (Kilgarriff et al., 2004) such as the Roget's Thesaurus of English Words and Phrases (Roget, 1911); Machine Readable Dictionaries (MRDs) such as the Oxford Dictionary of English (Stevenson, 2010) or ontologies such as the Omega Ontology (Philpot et al., 2005).

However, the most widely used knowledge base in the NLP field for WSD purposes is the lexical database WordNet (Fellbaum, 1998). WordNet is a huge network or graph of concepts (called *synsets*), each of them representing a set of synonyms related to the concept, thus each synset can be seen as a different sense of a word. This way, ambiguous words will present more than one synset in WordNet. Synsets are connected to each other through links representing semantic and lexical relations between them such as antonymy, homonymy, hyper and hyponymy, or meronymy. Although this database can be seen as a thesaurus since it groups similar words together, it has characteristics taken from ontologies, considering that it contains taxonomies as well as information about the semantic relations between concepts. A synset also contains a brief definition of the concept (called *gloss*), hence in some cases WordNet can also be considered as a MRD. Figure 2.2 shows an example of the WordNet structure, illustrating different concepts (synsets) and relations between them.

Several different techniques can be classified as knowledge-based WSD, but most surveys subclassify them into three categories:

- **Definition overlapping**: One of the most intuitive techniques for performing knowledge-based WSD takes the definitions or glosses of the possible senses of an ambiguous word into account. Although there exist many different variations of this technique, most of them are based on the Lesk algorithm (Lesk, 1986): in this method, the most

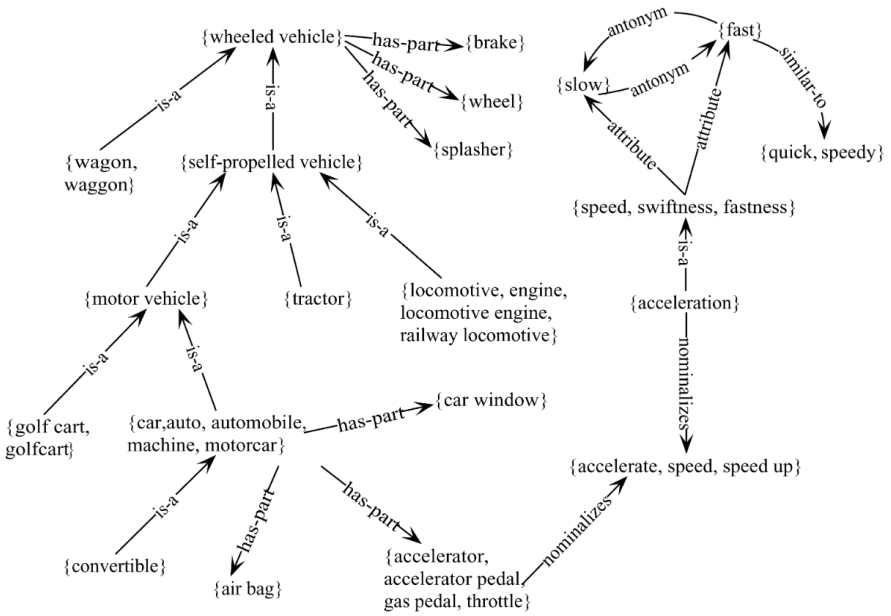


Figure 2.2: An excerpt of the WordNet semantic network (Navigli, 2009).

suitable sense for each word in a context is determined by computing the overlapping between every pair of senses s_1 and s_2, s_1 belonging to the set of senses of word w_1, and s_2 belonging to the set of senses of word w_2. The score assigned by this overlapping is maximized for every pair of words in the context, and hence a sense is assigned for each of the words. The computational drawbacks of this technique (the number of calculations exponentially increases with the size of the context) have been relieved by the development of alternative implementations of this method. For example, a simplified version of this algorithm assigns the most suitable sense for an ambiguous target term by selecting the sense with highest overlap between its definition and the context of the target word (Kilgarriff and Rosenzweig, 2000). Other variations of the algorithm extend the algorithm for giving more importance to *n-words* overlappings, that is, overlappings that contain more than a simple word. In the work by Banerjee and Pedersen (2003), concepts extracted from WordNet relations such as hyponyms or hyperonyms are also considered for computing the score for a pair of senses (synsets from WordNet) for developing an extended measure to compute semantic similarity, which can be used for performing WSD.

- **Selectional preferences**: This technique is one of the earliest method developed in knowledge-based WSD. The main idea behind it is the fact that specific linguistic elements usually receive arguments of a particular semantic class (Agirre and Martinez, 2001). If this idea is extrapolated to WSD, one can expect a specific sense of an ambiguous word to receive specific semantic arguments. Although these preferences or constraints can be learned by counting the number of times two words co-occur in

a corpus sharing a specific relation between them, the generalization of the technique to learn selectional preferences between a word and a semantic class, or between two semantic classes is a more complex task. Some other algorithms such as bayesian networks or hidden markov models have been explored for deriving these selectional preferences.

- **Semantic similarity measures**: The last category in which systems performing knowledge-based WSD can be classified relies in the definition of measures for computing the semantic similarity between two concepts, that is, measures which illustrate how close are two concepts (in this case, senses or synsets from WordNet) regarding to their semantic characteristics. There exist many different semantic similarity measures, from the earliest based on the shortest paths between senses in a taxonomy such as WordNet, either considering undirected (Leacock et al., 1998) or directed relations (Hirst and St-Onge, 1998). The concept of *information content* introduced in Resnik (1995) has been also widely used in semantic similarity studies. The information content of a concept measures its specificity by means of its probability of occurrence in a large corpus. According to this concept, the semantic similarity between two concepts is the information content of the first common ancestor in a given taxonomy (for example, WordNet). Another new concept called *conceptual density* is presented in Agirre and Rigau (1996). In this case, the definition of the measure itself is related to the context of the target word to be disambiguated, although information from a taxonomy is also required. The conceptual density of a concept C or specific sense is directly proportional to the number of senses in the hierarchy rooted by C, weighted with respect to the number of hyponyms that each sense presents, and inversely proportional to the total number of descendants of C. The most suitable sense of an ambiguous target word will be the one with highest conceptual density, given its context. Although this last measure includes the disambiguation of a given instance in the process of calculating it, in most of the cases explained above a process for performing WSD based on the similarity measure is needed. This process is usually based on the local context of the occurrence of a target word, that is, measures are computed for words in the vicinity of the target word, this way reducing the combinatorial explosion given by a high number of similarity computations. This process of analysing the local context of a target word is also known as the *one-sense-per-collocation* strategy (Yarowsky, 1993), which states that words next to the target word provide strong enough information to disambiguate it.

  A comparison between most of the previously mentioned semantic similarity measures for performing WSD is presented in (Pedersen et al., 2005), and finds the information content (IC)-based measure presented in (Jiang and Conrath, 1997) to be a good similarity measure for WSD. This measure takes into account the information content score of each concept in the considered pair, apart from the information content of their first common ancestor.

  Finally, a last subcategory of WSD processes which use similarity measures considers

the global context of the target ambiguous word for perfoming the disambiguation. The methods in this subcategory usually rely on the concept of *lexical chain*. A lexical chain is a sequence of related words in a text, such that the connection between two words in the chain is created by means of lexical cohesion, that is, cohesion related to a semantic relationship, such as "part-of" or "is-a" (Morris and Hirst, 1991). This semantic relationship can be extracted from a taxonomy such as WordNet. In the first works that make use of lexical chains, disambiguation is performed by considering those senses of the words in the text which can be easily added to an existing lexical chain, created with the senses of the previous words in the text. In an initial implementation of the algorithm, there exists a high risk of spreading mistakes toward subsequent disambiguation decisions, since a word is disambiguated by taking into account the already selected senses of the previous words. For avoiding this, all the possible interpretations of the words must be considered (Barzilay and Elhadad, 1999). Following this technique of considering all the possible meanings of the words in the text, multiple lexical chains can be merged into a graph (Galley and McKeown, 2003). This graph is then processed for performing the disambiguation itself, by computing the connections between all the occurrences of a target word and the senses related to it in each case. Hence, the same most suitable sense is assigned to all the occurrences of the target word in the considered context, in what is called the *one-sense-per-discourse* strategy (Gale et al., 1992b). This strategy is followed in the work by Mihalcea et al. (2004), in which a graph is also built by connecting all the possible senses in the text through their WordNet relations, but the final disambiguation step is performed through the PageRank algorithm (Brin and Page, 1998): The most suitable sense for all the occurrences of a target word will be the one with highest PageRank score.

As we will see in subsequent sections, the approach used for the first part of the thesis, focused on Cross-Lingual Word Sense Disambiguation, could be classified as unsupervised WSD, since it makes use of an unsupervised technique for creating an automatically generated statistical bilingual dictionary, which offers the possible word senses in the target language for each ambiguous word in the source one, and also an unsupervised algorithm for representing knowledge in a structured way and operating with that data structure (co-occurrence graph) for selecting the most suitable sense of an ambiguous word in each context. However, we have adapted our method for the second part of the thesis, Word Sense Disambiguation in the Biomedical Domain. In that case, we can consider our approach to be a knowledge-based technique, considering that we make use of an external ontology, which will be explained in Section 2.3, to annotate our documents, and hence we cannot further maintain that we do not make use of external resources. Nevertheless, our method is not supervised, given that we do not have information about the correct sense of each occurrence of a target term.

2.2 Multilinguality

In this Section we will explore the application of multilingual resources and techniques in different Natural Language Processing tasks, and we will focus on the main task of this thesis: Word Sense Disambiguation. Many NLP tasks can benefit from resources offering information in different languages, hence in this section we will analyse the nature and availability of some of those resources. Regarding the application of multilinguality to WSD, we will briefly define the Cross-Lingual Word Sense Disambiguation (CLWSD) task, which is the particular field of study that will be tackled in the first chapters of this thesis. Different systems addressing CLWSD will be presented and their characteristics compared, although a more thorough study of the results achieved by those systems in a particular CLWSD task will be presented later on, in Chapter 4.

2.2.1 Multilingual Resources in NLP and WSD tasks

As we mentioned in Section 2.1, Wordnet is probably the most important resource when it comes to the aim of gathering the lexical knowledge in a structured format. However, from the development of WordNet, many other projects have undertaken the task of enlarging this base of lexical knowledge in order to include information written in other languages. The development of multilingual databases similar to WordNet has been historically tackled through two different approaches: The EuroWordNet project (Vossen, 1998) is based on the independent construction of WordNet-like databases for every additional language involved in the project, apart from English (initially Dutch, Italian and Spanish, although other languages such as French, German, Czech or Estonian were added to the project in subsequent phases). In a second step of its development, synsets from different languages were linked together through equivalence relations, called Inter-Lingual Indexes (ILIs). This way, from a given synset in any language the user has access to similar synsets written in other languages of the database. The second approach was used in the development of the MultiWordNet project (Pianta et al., 2002). In that case, the construction of databases with synsets from other languages (initially Italian) is based on an "expansion model": the creation of new synsets is related to their correspondence with a "source" WordNet (typically the English WordNet). This way, semantic relations are imported from the original database, assuming that similar relations can be found between similar synsets in the new database. This paradigm has the advantage of being less complex and offering higher level of compatibility across the databases, although the dependency of the new databases with the original one in terms of structure is higher, as well as the probability of spreading mistakes along those new databases (Vossen, 1996).

One of the most important drawbacks in the development of those Wordnet-like multilingual databases is the fact that they are manually created. This leads to a high cost in implementation and maintenance, as well as a general trend to spend more efforts in resource-rich languages. The multilingual semantic network BabelNet (Navigli and Ponzetto, 2010) is the most important multilingual resource created as a solution to those drawbacks. The main idea behind BabelNet is the automatic mapping between synsets of WordNet and pages from Wikipedia. Through a disambiguation algorithm, a particular sense from a synset in WordNet is linked to the page of the English Wikipedia which describes that sense. This algorithm takes into account the contexts of both the WordNet sense and the Wikipedia page. The context of the WordNet sense is composed of its synonyms, hyper and hyponyms, sisters (synsets with common direct hypernyms) and lemmas from its gloss. The context of the Wikipedia page contains its sense labels, lemmas from the titles of pages linked to the target page, and categories. A link will be created between the WordNet sense and the Wikipedia page which maximizes the intersection of both contexts. Multilinguality is achieved through the inter-language links from the Wikipedia page. However, there exist cases in which Wikipedia does not offer translations of a page for some of the considered languages. In those cases, the BabelNet approach relies on the use of Machine Translation techniques, based on the SemCor corpus (Miller et al., 1993), for translating both sentences from that corpus in which the analysed sense occurs, and Wikipedia sentences with links to the page of interest. With this information, the most frequent translation in each language is added to the BabelNet network. Figure 2.3 shows the process of generating the BabelNet synsets from Wikipedia, WordNet and the SemCor corpus.

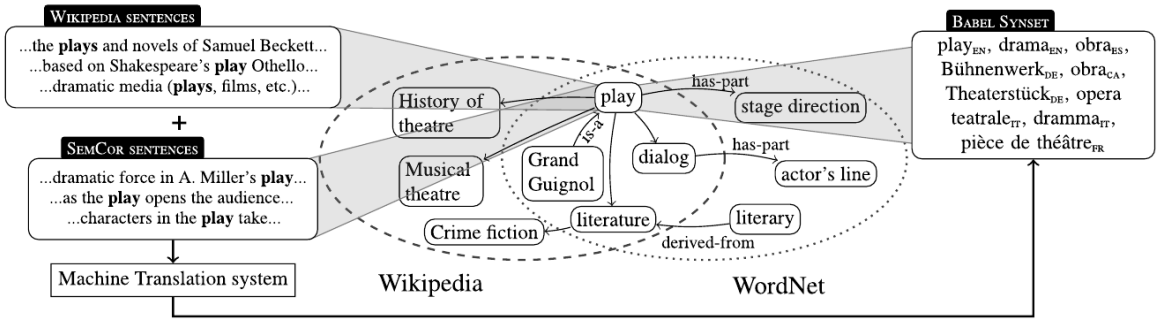


Figure 2.3: An illustrative overview of the creation of BabelNet synsets (Navigli and Ponzetto, 2010).

Apart from those structured representations of multilingual information, it is important to mention other resources which present a more relaxed, semi-structured or even unstructured representation of the information under a multilingual point of view. These resources are parallel and comparable corpora. Although we can find many definitions of those concepts and how they are created, for the purposes of our work, we will consider parallel corpora as collections of texts written in different languages and aligned to each other at some level. For example, a sentence-aligned parallel corpora is composed of two or more sets of sentences (written in two or more different languages, respectively, and distributed among a number of documents), so that for every sentence S_i^1 written in a language L^1, its exact translation S_i^2 can be found in a second language L^2. As we will see in subsequent chapters, parallel

corpora, and more particularly sentence-aligned corpora, are normally used for building statistical bilingual dictionaries through the use of tools such as GIZA++ (Och and Ney, 2003). A comparable corpora, however, will contain documents written in two or more different languages so that we can find a correspondence between document D_i^1 written in language L^1 and document D_i^2 written in language L^2, although that correspondence is merely topic-related, that is, the second document is a translation of the first one with regard to the topic and main meaning of the text, although no direct translations are established between sentences.

Some of the most commonly used parallel corpora are the Europarl corpus (Koehn, 2005), containing proceedings from the European Parliament, or the JCR-Aquis corpus (Steinberger et al., 2006), containing mostly legal documents from different sources of the European Union. Both of them are sentence-aligned and present documents written in more than 20 languages (all the official languages spoken within the European Union).

There exist a close relationship between semi-structured resources such as parallel corpora and structured resources like WordNet or BabelNet, since those kinds of corpora can usually be used for building semantic networks. For example, Kazakov and Shahid (2010) performs a word alignment over the sentence-aligned Europarl corpus through the use of the GIZA++ statistical aligner, which will be described in Chapter 5, and presents the notion of multilingual synsets, as sets of translations from a target word written in a "pivotal language", in this case English. These sets of "multilingual synonyms" are then post-processed and merged for reducing redundancy, through a process of detecting close words in terms of edit distance, and also detecting synonyms between words written in the same language.

Parallel corpora have been widely used as a source of information to perform WSD. One of the first analysis of their potentiality for disambiguation was presented in (Resnik and Yarowsky, 1999), in which an evaluation frame and an approach for measuring the distance between senses were proposed. Diab and Resnik (2002) proposed a method for automatically tagging senses in big parallel corpora, based on the use of sense inventories for each of the languages in the corpus. A supervised model that uses multilingual features for training a classifier was presented in (Banea and Mihalcea, 2011). These features are extracted by translating the context of the ambiguous words to different languages. The abovementioned resource generated by Kazakov and Shahid (2010) is then used for performing Word and Phrase Sense Disambiguation and analysing the reduction of lexical ambiguity of English words (Kazakov and Shahid, 2013).

We can also find works in the literature that address monolingual Word Sense Disambiguation tasks through the use of multilingual knowledge: for example, an unsupervised approach that uses a dictionary with definitions of the different senses of ambiguous words as the only available information is proposed in the work by Fernandez-Ordonez et al. (2012). This technique applies a variant of Lesk's algorithm for identifying the combination of senses that maximize the overlapping between their definitions, given a sequence of words. Senses from an ambiguous target word are extracted from WordNet, and all the sense definitions are automatically translated through the Google Translate API. Given a particular context for the

target word, the sentence is also automatically translated into all the considered languages (in this case, French, German and Spanish). The simplified Lesk algorithm is then separately applied over each of the additional languages, and the most suitable sense in each language is selected, depending on the overlapping between the translated context and the translated sense definitions. Finally, the final sense in English is determined using a voting system based on the multilingual senses already selected.

The study of monolingual WSD in different languages, and the combination of the results obtained by separate classifiers for each language for performing multilingual WSD is presented in the work by Dandala et al. (2013a). The main source of knowledge in this case is Wikipedia, which provides information in many different languages, as well as links between those languages, called interlingual links. In a first step, sense annotated corpora are generated in English, German, Italian and Spanish from the information extracted from the Wikipedia pages, using an approach described in previous works (Dandala et al., 2013b). Basically, paragraphs containing a particular ambiguous target word, and being part of a link or a piped link in Wikipedia are selected. Every possible label for representing a sense of the target word in every case is extracted from the leftmost component of the links. Finally, a manual clustering algorithm based on heuristics is applied for grouping the labels and creating the possible senses of the target word. Once that the sense annotated corpora are created, a Naïve Bayes machine learning system using both local and global context features is run for performing monolingual WSD over the proposed languages. The next step is to analyse the possibility of performing automatic multilingual WSD for all the languages at the same time, and comparing the results with the monolingual results achieved using the mentioned system. For this purpose, two different ways of translating paragraphs from a source language are proposed: the first one is based on the Google Translate API, and the second one is based on the interlingual links from Wikipedia. Using any of those two approaches, every considered language is taken as source language (called reference language), and translated into the remaining ones (called supporting languages). The same local and global features are extracted from the multilingual information, and the Naïve Bayes algorithm is used for performing the final disambiguation in each of the reference languages. The use of multilingual features offers an overall improvement of the results obtained for each of the monolingual experiments conducted in the first phase of the study.

2.2.2 Cross-Lingual Word Sense Disambiguation

Cross-Lingual Word Sense Disambiguation (CLWSD), which will be described in detail in Chapter 4, aims to find the most suitable translation for a given ambiguous term written in a source language, typically English, in a target one. The motivation of the CLWSD task, which is a particular case of the WSD problem itself, comes from the scarcity of sense inventories and sense-tagged corpora, and the interest in evaluating the performance of WSD systems in real problems. SemEval competitions represent some of the most important

venues for addressing many different NLP tasks. In particular we focused on task 3 of the SemEval 2010 competition (Lefever and Hoste, 2010a) and task 10 of the SemEval 2013 competition (Lefever and Hoste, 2013), since those are the tasks devoted to CLWSD in the past few years, so they represent a good indicator on the quality and accuracy of state-of-the-art systems addressing the same problem. In these tasks, the Europarl corpus was proposed as the main knowledge source, especially for unsupervised systems which rely on a source of knowledge for building the data structures that will be used for performing the disambiguation.

The UvT-WSD system (van Gompel, 2010) is the approach that obtained the best results in the 2010 competition. It works on a basis of local and global context features, following the idea behind typical supervised approaches as we described in section 2.1.2. Local features, extracted as usual from the neighbourhood of the target term, are related to the words themselves, as well as their part-of-speech tags and their lemmas. Global features are represented by a bag of keywords, and the method for extracting these keywords is based on the system proposed by Ng and Lee (1996): the conditional probability of occurrence of a sense s of the target word w given a keyword k is calculated by dividing the number of co-occurrences of w presenting sense s and keyword k, by the total number of sentences in which w co-occurs with k, having w any sense. Thresholds are defined for establishing the minimum value of this conditional probability, as well as the minimum and maximum number of keywords in the bag of words. Once those features are calculated, after a voting phase in which different classifiers are built depending on a feature selection process in which only some of the described features are used, the Machine Learning algorithm used for performing the disambiguation is IB1, an implementation of the *k-NN* algorithm. Through this implementation, a particular classifier, denoted as "word expert" is built for each of the target words in the test dataset. In the SemEval 2013 competition, the system called WSD2 (van Gompel and van den Bosch, 2013), which is a new version of UvT-WSD with an improved hyperparameter optimisation phase for each of the word experts, was also able to obtain the best results among the supervised systems.

An alternative supervised approach makes use of a Naive Bayes classifier for performing the disambiguation (Vilariño et al., 2010). In this technique, weighted probabilities for the possible translations of the target word are extracted through the GIZA++ tool, or using the API of the Google translator, depending on the version of the algorithm. The Naive Bayes classifier is then used for computing the probability of the target word w to be related to each of its possible translations t_i, given the context words of the sentence in the test instance.

The HLDTI system (Rudnick et al., 2013) is another supervised system which participated in the 2013 competition. The system uses Maximum Entropy classifiers divided in three different layers for performing the disambiguation, although each layer itself is able to offer results for the task. Features used for classification are divided into target word features (POS tag, literal form and lemma), unigram features from a window of 3 words (literal form, POS tag, lemma and word+POS tag), and bigram features from a window of 5 words (bigrams and bigrams+POS tags). With this information, the first layer, considered monolingual,

trains a classifier for each of the target words. In a second layer, multilingual information represented by the correct translations of the target word in other languages is introduced, either directly at training time, or estimated by the first layer at testing time. The last layer is a Markov network, also called Markov Random Field (MRF) which only makes use of the classifiers designed in the first layer, for finding the most suitable translation for all the target languages jointly.

Regarding unsupervised approaches, the T3-COLEUR system (Guo and Diab, 2010) offers the best results in the 2010 competition. It is based on the extraction of probability tables from the Europarl corpus, also through the GIZA++ tool. The main assumption considered for building the system is that every sense of a target ambiguous word in the English language is related to a particular word in the target language. Hence, the CLWSD process is based on two different modules. The first module is a monolingual WSD graph-based technique previously developed by the authors (Guo and Diab, 2009). The second module is based on the extraction of probability tables using the GIZA++ statistical aligner, although assigning a specific sense to the English word of the alignment. This way, each of the possible senses of an ambiguous word in English (extracted from WordNet) is related to a particular word of the target language. When a test instance is considered, the first step will be to determine the monolingual sense of the English target word, through the WSD algorithm. After that, a look-up to the probability table will provide the most suitable translation for that English sense in the target language.

The UHD system (Silberer and Ponzetto, 2010) is somehow similar to the approach described in this thesis, considering that it also builds a co-occurrence graph based on the aligned contexts of the target word. However, the main idea of the technique is based on the HyperLex algorithm (Veronis, 2004), already described in Section 2.1.3. In the work by Silberer and Ponzetto (2010), a monolingual graph is first created for each ambiguous target word in the original language (typically English). The nodes of the graphs are words (nouns and adjectives) co-occurring with the target word in different contexts, and the weights between nodes are calculated based on the probabilities of co-occurrence. In a second step, a multilingual graph is generated by expanding this monolingual graph with information extracted from the aligned contexts of the target word found in the multilingual corpus (in this case, Europarl). New co-occurrence links are generated between words written in the new language, and additional translation edges are created for linking words written in different languages. Those edges represent possible translations of the words co-occurring with the target word. Once the multilingual graph is created, the PageRank algorithm is run for finding the nodes (called hubs) which represent the senses of the target word, in the English language. Then, the Minimum Spanning Tree (MST) of the graph is computed. The root of this MST will be the target term, and the hubs found in the previous step will be directly connected to it. From this MST the disambiguation of a test instance can be performed by selecting the hub which is more related to words in the context. A subgraph only containing nodes related to this selected hub will be extracted from the main multilingual graph, and the possible translations of the target word will be found in that subgraph and ranked according to their translation counts.

Other unsupervised approaches have been applied for tackling the CLWSD problem: for example, the system LIMSI (Apidianaki, 2013) addresses the task by using vectors of features extracted from the corpus. The main background behind this system is the algorithm for semantic clustering introduced in previous works (Apidianaki, 2008, 2009). After extracting a word alignment through GIZA++, a vector of features is generated. This vector is composed of words that co-occur with the target word whenever it is translated to a particular word in the target language. Each feature receives a total weight which depends on both its global weight, related to its co-occurrence with any of the translations of the target word, and its local weight, related to its co-occurrence with a particular translation. Vectors of features for each possible translation (called translation vectors) are then grouped into clusters containing only those semantically related translations of the target word. Those clusters are then used for performing the disambiguation of a test instance. Although it is an unsupervised approach, for performing the disambiguation in the French language it uses knowledge provided by an external resource apart from the Europarl corpus, which is the abovementioned JRC-Acquis corpus.

Topic modelling is another technique explored by some of the unsupervised systems participating in the SemEval competitions, such as XLING (Tan and Bond, 2013). In this technique, topic models are generated in a training step from the source corpus using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The main hypothesis is that the different senses of a target word will be classified into different topics by the LDA algorithm. This way, when a new test instance is received, it is matched against the topics generated in the training step and classified as the most probable topic. Then, sentences from the training phase that belong to that topic are ranked depending on their similarity with the test instance. The final step provides translations of the target word from those sentences ranked in the previous step, using a word alignment of the Europarl corpus created with GIZA++.

The NRC-SMT system (Carpuat, 2013) uses a phrase-based statistical machine translation system, extracting knowledge only from the Europarl corpus in its first run, and adding information from news data in a second run of the system. The disambiguation decisions for each of the test instances are based on the local source context (source phrases of different lengths, from the translation tables of the machine translation system), and on the local target context, represented by a 5-gram language model written in the target language.

Finally, we can find other systems that did not participate in any of the competitions, although they have addressed the tasks proposed in them: the ParaSense system (Lefever et al., 2011) is a supervised, memory-based algorithm that builds different classifiers using both local context features and binary bag-of-words features. In this case, the local context features are related to information about the full form, lemma, POS tag and chunk of a window of three words before and after the ambiguous target term. Unsupervised systems as the multilingual system described in (Navigli and Ponzetto, 2012) also addressed the problem without participating in the competitions. This system exploits the multilingual knowledge base BabelNet (Navigli and Ponzetto, 2010), for performing WSD and CLWSD, obtaining very competitive results. From test instances, subgraphs from BabelNet containing words

from the sentence (including the target word) and its translations are extracted. The different languages involved in the dissambiguation task are considered at the same time in the extracted sugraph, and hence the disambiguation is performed jointly for all the involved languages. Both works make use of the GIZA++ tool, the first one as a main aligner for extracting a bilingual dictionary, and the second one for proposing the most frequent sense translations when no sense assignment is attempted.

2.3 WSD in the Biomedical Domain

As we mentioned in Chapter 1, the motivation for addressing the second part of this thesis, WSD in the biomedical domain, is based on the analysis of the possibilities of applying multilingual techniques such as those described in Section 2.2 to a particular domain which presents many cases of ambiguity and thus many opportunities for implementing WSD techniques which improve the accuracy and performance of NLP systems. These WSD techniques represent a key step to automatically access, retrieve and process the increasing amount of available unstructured textual information (Savova et al., 2008) in the field, and should be implemented in almost every system attempting to perform more complex NLP tasks, such as summarization (Plaza et al., 2012), information extraction (Chai and Biermann, 1999), or Literature Based Discovery (Preiss and Stevenson, 2015).

However, there are also multiple scenarios in the biomedical domain in which data scarceness is one of the major issues for building a system that successfully performs Natural Language Processing (NLP) tasks. These scenarios include studies developed in low-income or middle-income countries in which health research efforts and resources are unequally distributed (Røttingen et al., 2013), or works regarding low resource languages. Also, information related to specific tasks such as the study of rare diseases is scarce and difficult to summarise, as well as time-consuming for the few experts in the area (Aymé, 2016). In those cases we are talking about poorly documented problems, for which most of the available corpora in the literature will be small (Valmaseda et al., 2016). Hence, our goal when addressing WSD in the biomedical domain under a multilingual perspective is to determine whether the inclusion of multilinguality within WSD techniques allows us to improve monolingual systems performing WSD, but also to analyse under which conditions these multilingual techniques are particularly useful, regarding the amount of available information.

The dominant knowledge source in the biomedical domain, in analogy to WordNet when it comes to general NLP, is the Unified Medical Language System (UMLS) Metathesaurus (Humphreys et al., 1998), which attempts to gather all the knowledge about the biomedical domain in an ontology containing all the available resources in the field. UMLS presents a structure with references and information from diverse sources and vocabularies with

medical terminology: MeSH with information from biomedical literature[i], SNOMED from clinical repositories (Stearns et al., 2001) or genetic knowledge bases such as GO (Ashburner et al., 2000) or OMIM (Hamosh et al., 2005). Figure 2.4 represents the different sources from which the information that populates the UMLS database is extracted. From those resources, the database is organized in a four-level metathesaurus in which the first and most important layer is built around Concept Unique Identifiers or CUIs, which are codes that unequivocally identify medical concepts inside it. The rest of the levels contain Lexical Unique Identifiers (LUIs), with lexical variants of CUIs, String Unique Identifiers (SUIs), with descriptive chains containing the possible forms in which the concept can be found in a biomedical text, and Atom Unique Identifiers (AUIs), which correspond to an occurrence of each SUIs from a particular resource or vocabulary in the metathesaurus. For the purposes of this thesis, we will focus on the concepts that can be found inside a biomedical text, which are represented by their CUIs in the UMLS database. As we will explain in subsequent chapters, the set of possible CUIs to which an ambiguous biomedical term may refer represent the ambiguity of the problem we need to solve. UMLS also presents a semantic network in its structure, which provides a consistent classification of the concepts and relations between them, organized through semantic types (high level categories comprising a set of concepts) and semantic relations such as "is-a", "associated-with", "part-of", "treats", "diagnoses", etc. Hence, CUIs in the UMLS database are linked to other CUIs depending on the different relations between them (Chasin et al., 2014).

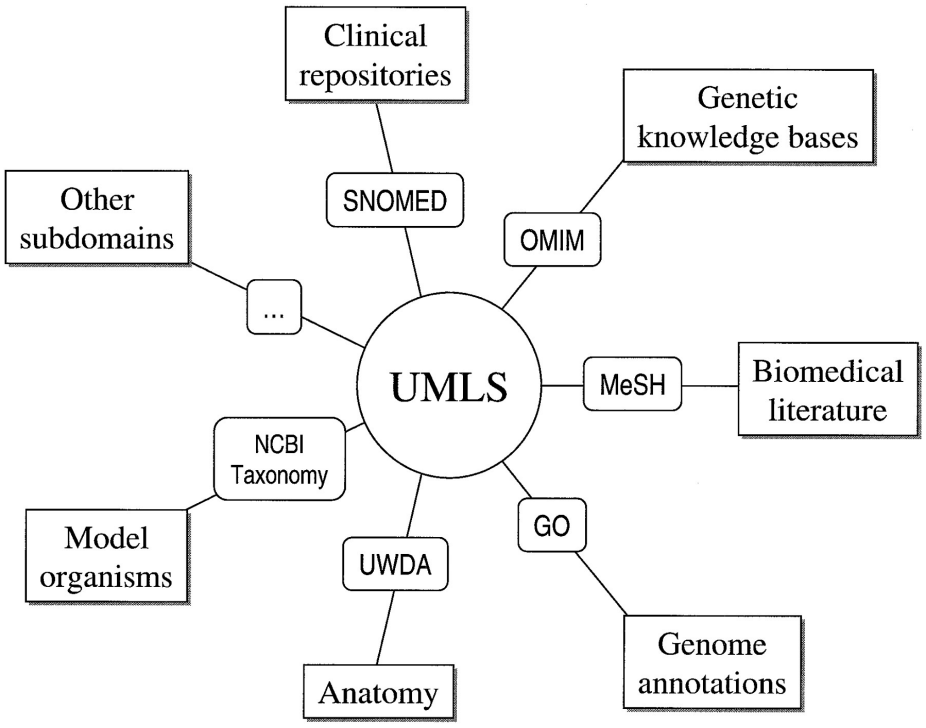


Figure 2.4: The various subdomains integrated in the UMLS (Bodenreider, 2004).

A very important resource regarding the UMLS database is the MetaMap program (Aronson,

[i]https://www.nlm.nih.gov/mesh/

2001), a tool that splits the text inside a biomedical document into phrases, and maps each of those phrases onto a set of UMLS CUIs. The main process of the MetaMap program, which will be described in more detail in Chapters 6 and 7, includes tagging and parsing the target text using the Xerox part-of-speech tagger (Cutting et al., 1992) and the SPECIALIST lexicon (McCray et al., 1994). After this step, the program generates variants for each given phrase, including acronyms and abbreviations, synonyms, as well as derivational, inflectional and spelling variants. Then, a candidate set of mappings is generated and evaluated for proposing the most suitable mappings of the analysed phrase.

It is commonly accepted that most of the systems performing biomedical WSD can be separated into two main groups: data-driven or algorithms that need labelled training data, and knowledge-based techniques (Agirre and Edmonds, 2007b; Schuemie et al., 2005). The first category, also called supervised techniques, usually applies machine learning (ML) algorithms to labelled data to develop a model, based on features extracted from the context of the ambiguous words. The development of these features requires a comprehensive understanding of the problem being addressed (Moon et al., 2012). Many works performing supervised WSD can be found in the literature, most of them making use of these linguistic features that are usually employed for performing WSD in more general domains (Stevenson and Guo, 2010a). Features such as part-of-speech (POS) tags, unigrams and bigrams are used by Joshi et al. (2006) for training Naïve Bayes classifiers, decision trees and Support Vector Machines (SVMs) and their results are compared. SVMs are also used by Gaudan et al. (2005) for abbreviation disambiguation. Vector Symbolic Architectures (VSA) have been used by Moon et al. (2013) for encoding vector representations for the ambiguous term and each of its senses. This representation can be reversed for new instances containing the ambiguous term in order to recover the appropriate sense for the context. More recent works have also applied state-of-the-art deep learning techniques such as neural word embeddings to acronym disambiguation (Wu et al., 2015). In this work, different techniques are implemented for deriving word embeddings of ambiguous terms, and their performances are evaluated and compared inside a system which uses a SVM algorithm taking the word embeddings as inputs.

As in general purpose WSD, the bottleneck caused by the scarcity of labelled resources remains a major problem. Some semi-supervised works attempt to relieve this issue by introducing "pseudo-data" to the training examples (Stevenson and Guo, 2010b) or by creating automatically extracted and annotated training corpus (AEC) for building Machine Learning systems (Jimeno-Yepes and Aronson, 2010). In the first work, two different approaches are used for automatically generate labelled examples: in the first one, unambiguous lexicalizations (LUIs) of senses (CUIs) from an ambiguous target term are extracted from UMLS. In those cases in which no monosemous LUI is found in the database, relatives from the particular CUI are analysed in order to find those monosemous LUIs. With those monosemous lexical chains, examples of abstracts containing the lexical chains are annotated with that particular CUI. In the second approach, co-occurrence information stored in the UMLS database is accessed for extracting those CUIs that usually co-occur with a particular CUI of the ambiguous target word. Abstracts containing both the target word and as many

co-occurring CUIs as possible are likely to be labelled with that particular CUI. Once all the training examples have been generated, a supervised approach based on vector space models is implemented, with a range of local and global features for representing the vectors. The generation of these "pseudo-data" for their use in supervised approaches is easier when the ambiguity that needs to be solved is introduced by the possible extended forms of abbreviations or acronyms. In that case, the known extended form of an abbreviation can be used for finding abstracts containing it. Then, the string representing the extended form can be contracted to its acronym, and the abstract labelled as belonging to the sense represented by the extended form (Xu et al., 2012).

Regarding knowledge-base methods, they are usually based on initially untagged corpora, and rely on external resources such as UMLS. Graph-based techniques such as the one proposed in this thesis have been usually explored in the literature. In the work by Agirre et al. (2010), the UMLS database is directly converted into a graph which will be used for performing the disambiguation process. CUIs from the UMLS database become the nodes of the graph, while information from tables representing semantic relations (the *MRREL* table) and co-occurrence information (the *MRCOC* table) is used for generating the edges. Once the graph is created, the disambiguation algorithm that is selected for performing the final disambiguation is the Personalized PageRank (Agirre et al., 2013), which will be detailed in subsequent chapters of this thesis. Through this algorithm, given an ambiguous term and its context, after an annotation process through the MetaMap program, the most relevant CUIs according to the context are highlighted in the graph, so they can be ranked for selecting the most suitable sense (CUI) for the ambiguous target term. Given its similarity to the approach described in this thesis, we are particularly interested in comparing the results obtained by our co-occurrence graph-based technique with this approach, in different test datasets.

Two different knowledge-based systems to which we will also compare our proposed technique is the MRD (McInnes, 2008) and 2MRD (McInnes et al., 2011) techniques. The first one creates an instance vector containing words in the context of the ambiguous target term, in a test instance. This vector is then compared to concept vectors previously created for each of the possible senses of the target word, using cosine similarity, and the most similar one is selected to be the most appropriate sense for the test instance. Different features are explored for creating the concept vectors: words in the definition of the particular CUI, words in the definition of the semantic type to which the CUI belongs, words in both CUI and semantic type definitions, and finally words in the CUI definition for those that present it, and words in the semantic type definition otherwise. This approach is evolved in the work by McInnes et al. (2011) for performing acronym disambiguation. Extended definitions of each long or extended form of the acronyms are found in the UMLS database, by taking words from the definition of the extended form itself, if available, and from the definitions of concepts that present parent/children and narrow/broader relations with the target CUI. Once these extended definitions are created, a second order vector is created by calculating co-occurrences between the extended definition of each long form of an acronym, and words in abstracts which present that long form. A co-occurrence matrix is created for each long form, containing words from the extended definition in the rows, and words from abstracts

containing that long form in the columns, and calculating the log-likelihood ratio of the co-occurrence. A centroid vector is calculated from those vectors representing words from the extended definitions (rows of the matrix), which will represent the second-order vector of the long form. For each test instance containing an ambiguous acronym, a second order vector (instance vector) is created in a similar way, although this time words in the rows of the co-occurrence matrix are context words of the test instances, and words in the columns will be those appearing in abstracts containing the ambiguous acronym. The second order vector, among those that represent long forms of the acronym, that presents higher cosine similarity to the instance vector will be selected as the most appropriate long form for the ambiguous acronym. This technique is applied by Jimeno-Yepes et al. (2011) for performing WSD over a different test dataset, composed by general words apart from acronyms (the extended definition is taken from the CUI representing each possible sense of a target word). We can find a similar approach in the literature that also transforms the knowledge contained in the UMLS database into a vector representation of the different possible senses of an ambiguous term (Tulkens et al., 2016). However, in this work the vectors representing the different senses are extracted through techniques based on word embeddings: each possible sense of an ambiguous term is represented by a concept vector created by combining the word embeddings of all the words in its UMLS definition. For each test instance, a similar vector is generated from the word embeddings of words surrounding the occurrence of the ambiguous term, and the cosine similarities between that vector and the concept vectors of the possible senses are calculated for finding the most suitable sense.

Finally, the JDI system (Humphrey et al., 2006) assigns a specific semantic type for the ambiguous target term of a test instance, based on semantic type tables (ST-tables) built through a similar vector based approach. However, this method is restricted to ambiguous terms for which each of its possible senses belongs to a different semantic type, which restricts the overall performance and possibility of application of the technique.

Results achieved by the technique that we propose and explore in this thesis will be compared against those obtained by the works previously detailed, and more particularly against knowledge-based systems, making use of the UMLS database but avoiding the use of training labelled instances. Regarding multilinguality, as far as we have been able to find in the literature, the use of multilingual information for performing WSD in the biomedical domain has not been explored, and hence the main motivation of this part of the thesis will be to analyse this field of study and offer results either encouraging or advising against the use of multilingual information.

3

THEORETICAL BACKGROUND

It is not the beauty of a building you should look at; it's the construction of the foundation that will stand the test of time.

David Allan Coe

Contents

3.1 Introduction

This chapter presents the description of the main theoretical background which is the basis for all the techniques and algorithms that will be used in subsequent chapters, for addressing different Word Sense Disambiguation tasks under various perspectives. As we have stated in previous chapters, our technique is based on co-occurrence graphs. The only resource that we need for building these graphs is a corpus formed by a number of text documents. These documents contain the unstructured information which will be transformed into a structured knowledge base (the co-occurrence graph). In turn, this knowledge base will be the structured resource on which all the disambiguation techniques presented in this thesis will rely. The original documents used as source of information for building the graph will always be the same, that is, raw text documents extracted from different sources, depending on the task we are tackling. However, the nature of the elements that will eventually become nodes of the co-occurrence graph will change, in accordance with the requirements of each task.

3.2 Co-Occurrence Graph

We consider our source of information to be a corpus composed of a set of N documents, $S = \{D_1, D_2, ..., D_n\}$. Each of those documents is formed by a number of concepts, $D_i = \{c_1, c_2, ..., c_m\}$, such that the concept c_i can be found in n_i number of documents from the corpus. As we considered in Chapter 1, the appearance of a concept in a document is likely to be related to the general sense of the document, but not necessarily. Hence, in order to determine whether the co-occurrence of two concepts in the same document is statistically significant, we need to define a null model for representing co-occurrences which can be considered as happening by pure chance. In this null model, concepts would be randomly and independently distributed among a set of documents, and the probability of two concepts co-occurring by pure chance is calculated. We then compare the actual co-occurrences of each pair of concepts against this null model and select those that present a high statistical significance, that is, a low probability of being generated by the null model. In other words, if a co-occurrence can be easily (i.e., with high probability) generated by the null model, then it is not considered to be statistically significant. More specifically, a p-value p is calculated for the co-occurrence of two concepts inside the null model. If p lies below a given threshold next to 0, $p \ll 1$, then the appearance of the two concepts in a document is significant, and they are probably related to the same sense, according to our hypothesis.

For illustrating the process of determining significance of a co-occurrence, let us assume two concepts c_1 and c_2 appearing in n_1 and n_2 number of documents from S respectively. In order

to calculate in how many ways those two concepts could co-occur in exactly k documents, we can divide the complete set S in four different types of documents: k documents containing both concepts c_1 and c_2 at the same time, $n_1 - k$ documents containing only concept c_1, $n_2 - k$ documents containing only concept c_2, and $N - n_1 - n_2 + k$ documents containing neither c_1 nor c_2, given that N is the total number of documents in S. The number of possible combinations of co-occurrence is given by the multinomial coefficient:

$$\binom{N}{k, n_1 - k, n_2 - k} = \binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k} \tag{3.1}$$

Then, given two concepts randomly and independently distributed among N documents, and appearing in n_1 and n_2 documents respectively, the probability of those concepts co-occurring in exactly k documents, that is, the probability of co-occurring by pure chance, is given by:

$$p(k) = \frac{\binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k}}{\binom{N}{n_1}\binom{N}{n_2}} \tag{3.2}$$

if $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$, and zero otherwise.

Equation (3.2) can be rewritten in order to get an equivalent expression that is computationally easier to deal with. For this purpose, we introduce the notation $(a)_b \equiv a(a-1)...(a-b+1)$, for any $a \geq b$, and without loss of generality, we assume that the first concept, c_1, is the most frequent one, that is, $n_1 \geq n_2 \geq k$. Then:

$$p(k) = \frac{(n_1)_k (n_2)_k (N-n_1)_{n_2-k}}{(N)_{n_2}(k)_k} = \frac{(n_1)_k (n_2)_k (N-n_1)_{n_2-k}}{(N)_{n_2-k}(N-n_2+k)_k (k)_k} \tag{3.3}$$

where in the second form, the identity $(a)_b = (a)_c(a-c)_{b-c}$, valid for $a \geq b \geq c$. Finally, equation (3.3) can be rewritten as:

$$p(k) = \prod_{j=0}^{n_2-k-1}\left(1 - \frac{n_1}{N-j}\right) \times \prod_{j=0}^{k-1}\frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)} \tag{3.4}$$

This allows us to define the following p-value p for the co-occurrence of two concepts:

$$p = \sum_{k \geq r} p(k), \tag{3.5}$$

where r is the number of documents of our actual in which we can find c_1 and c_2 together. As we stated before, if p lies below a specific threshold p_0 next to 0, the co-occurrence is considered to be statistically significant and a link between c_1 and c_2 is created in the graph. The maximum p-value for statistical significance purposes usually found in the literature

is $p_0 = 0.01$ or $p_0 = 0.05$ (99% or 95% of statistical trust). We will use these maximum values for all the experiments conducted in the different WSD tasks that we will tackle. However, p_0 can also be considered as a threshold which determines the restrictiveness of the technique for linking two concepts in the resulting co-occurrence graph: when the graph is constructed, p_0 will indicate the highest value of p for which the number of co-occurrences of two concepts is considered to be statistically significant. As this threshold decreases, the graph will become more restrictive, and hence the number of edges in the graph will also decrease. In short, smaller values of p_0 lead to smaller, and hence more easily manageable, graphs. The influence of the value assigned to threshold p_0 will be deeply analysed in many of the experiments proposed in this work.

Finally, we have developed a way to measure the importance of a link between two concepts of the graph, considering the significance of their co-occurrence. Given equation (3.5), the weight of the link between two nodes i and j can be quantified in a practical way by defining it as $w_{ij} = \log(p_0/p_{ij})$, where p_0 is the selected threshold for the co-occurrence graph and p_{ij} is the p-value calculated using equation (3.5) and defining r as the actual number of co-occurrences between concepts c_i and c_j. This way, the weight of the link will be proportional to the order-of-magnitude difference between p and p_0.

It is important to notice that the approach described here has the advantage that it does not assume that concept frequencies are normally distributed, unlike some alternative measures of lexical co-occurrence (Hitchcock, 2009). For example, a chi-squared method would assume data to follow a Gaussian distribution, which is not valid for many cases, especially when the number of co-occurrences is small. Our data correspond to a particular distribution which would only approximates Gaussian for very large values, so chi-squared would not be recommended in this case. Hence we directly calculate how our actual data deviate from our distribution (null model). However, as we will see in the following sections, we have decided to compare our method with some of those alternative measures of co-occurrence, in order to empirically prove that our technique is able to obtain more reliable results for the kind of tasks that will be addressed in this thesis.

3.3 Comparison with other Methods

Regarding the statistical process described in the previous section, we are interested in comparing the model with other methods found in the literature for calculating the statistical significance of co-occurrences within a corpus. As we stated before, our method is based on the assumption that the considered data (co-occurrences at document level in textual corpora) correspond to a particular distribution able to better represent both low and high co-occurrence values. This way, we assure that the statistical model built in Section 3.2 is reliable and well adjusted to our data, independently of the size of the corpus we use in each case. In this section, we will perform the comparison between our technique and

the alternative methods by applying them to a particular Word Sense Induction (WSI) task, which will be described in Section 3.3.3. This task requires the creation of clusters of words for representing senses of particular ambiguous words, and hence a post-processing will be needed once that we have created the co-occurrence graphs (with either our technique or any of the alternative methods). In that post-processing, a clustering step will be performed over the nodes of the graph, for creating communities, that is, densely connected subgraphs that will represent the different senses of the ambiguous words. The community extraction step will be detailed in Section 3.3.2.

3.3.1 Alternative Methods

Although different alternative methods can be found in the literature for calculating the significance of lexical co-occurrences, we will focus on two widely used methods: the Chi-Square test (Fisher, 1925; Pearson, 1900) and the G-Test, initially developed by Fisher (1929) and introduced to computational linguistics by Dunning (1993). Both tests can be used for attempting rejection of the null hypothesis of independence of the proposed data. That is, for the purposes of this work, we can use the Chi-Square test and the G-Test in a similar way presented in Section 3.2, in order to propose a null model in which we calculate the probability of random co-occurrences (independence of data). Following the notation introduced in that section, we extract from the corpus, for each pair of concepts c_1 and c_2, the exact values of occurrence of each concept, n_1 and n_2 respectively, as well as the exact number of co-occurrences of both concepts inside the same document, k, which in this case we will denote as n_{12}. Through the calculation of the Chi-Square value χ^2 and the G-Test value G for the exact number of observed occurrences and co-occurrences, we can derive a p-value p which will determine whether we can reject (p-value next to 0) the assumption of independence proposed by the null hypothesis, and hence consider that the co-occurrence is not happening by chance and a link must be created between both concepts in a co-occurrence graph.

For the application of both the Chi-Square test and the G-Test, we need to obtain some coefficients based on the characteristics of the considered problem. In our case, we will denote the probability of occurrence of a concept c_i, occurring in n_i documents within a corpus containing N documents, to be

$$p_i = \frac{n_i}{N}. \tag{3.6}$$

The probability of co-occurrence of two concepts c_1 and c_2, appearing in exactly n_1 and n_2 documents and co-occurring in n_{12} documents, will be

$$x = \frac{n_{12}}{N}. \tag{3.7}$$

We call this value x since that will be the probability that correspond to the observed value (actual number of co-occurrences in the corpus).

The calculation of the proposed alternative methods relies on the following parameters:

- **Categories and degrees of freedom**: For the application of the mentioned tests, we need to define the number of possible categories into which a observation can fall. In our case, the possible categories into which a simple observation (a specific document in the corpus) of a particular pair of concepts c_1 and c_2 can be classify is whether they co-occur or not in the given document. Hence, the number of categories will be $k = 2$, and by definition, the number of degrees of freedom for the application of the Chi-Square test and the G-Test will be $d = k - 1$, that is $d = 1$.
- **Observed values**: Considering the proposed categories, the observed value for our problem, denoted as O_i, for a pair of concepts c_1 and c_2, is their actual number of co-occurrences, n_{12} for the category of "co-occurrence", and $N - n_{12}$ for the category of "non-co-occurrence".
- **Expected values**: Considering that the coefficient which will be used for both accepting or rejecting the null hypothesis is the actual number of co-occurrences, we need to define the expected value, E_i, for the defined categories. For the "co-occurrence" category, this expected value of total co-occurrences given the total number of documents N can be seen as the probability of occurrence of the first concept, p_1, times the probability of occurrence of the second concept p_2 (since the null model considers them to be independent), times this total number of documents N. Hence the expected value will be equal to $p_1 p_2 N$, and using equation (3.6) for each p_i, we have $\frac{n_1 n_2}{N}$. The expected value for the category of "non-co-occurrence" will be $N - \frac{n_1 n_2}{N}$.

As we explained in Section 3.2, the proposed CO-Graph technique calculates the probability of the particular configuration of occurrences and co-occurrences between two concepts c_1 and c_2, that is, the values for n_1, n_2 and n_{12}, by comparing it with respect to all the possible reshufflings of the same configuration (given by the multinomial coefficient expressed in equation (3.1)). As we can observe in the abovementioned parameters, both alternative methods assume the a priori probability of co-occurrence, or expected value $E_i = \frac{n_1 n_2}{N}$ for the category of "co-occurrence". That is, the a priori probability is calculated from the observed frequencies n_1 and n_2. Although these methods are able to obtain good approximations of the statistical significance when the number of documents used for creating the co-occurrence graph is large, their behaviour usually differ when that number of documents is small.

3.3.1.1 Chi-Square

The Chi-Square test is related to normal distributions of the considered data. In our case, it is not proven that our data can be adjusted to that normal distribution, however, we are

interested in testing this method in order to empirically test how it compares to our technique for transforming the information within the corpus into a structured representation. We will use the Chi Square test to determine whether a particular number of co-occurrences between two concepts in a corpus can represent evidence of dependency between those two concepts, or on the contrary, those concepts can be considered as independent.

The definition of χ^2 for a given number of co-occurrences, n_{12} is as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{3.8}$$

where O_i is the observed and E_i the expected value for the category "co-occurrence" ($i = 0$) and "non-co-occurrence" ($i = 1$). After substituting those values by their values for each of the categories, previously defined, we have:

$$\chi^2 = \frac{\left(N_{12} - \frac{N_1 N_2}{N}\right)^2}{\frac{N_1 N_2}{N}} + \frac{\left(N - N_{12} - N + \frac{N_1 N_2}{N}\right)^2}{N - \frac{N_1 N_2}{N}}. \tag{3.9}$$

Now we are able to use the definitions of p_i and x from equations (3.6) and (3.7), for obtaining the following formula:

$$\chi^2 = N\frac{(x - p_1 p_2)^2}{p_1 p_2} + N\frac{(x - p_1 p_2)^2}{1 - p_1 p_2}, \tag{3.10}$$

and, simplifying:

$$\chi^2 = N\frac{(x - p_1 p_2)^2}{p_1 p_2 (1 - p_1 p_2)} \tag{3.11}$$

3.3.1.2 G-Test

The G-Test is a statistical test related to the concept of likelihood ratio, that is, a comparison between a null model and an alternative model which offers an estimation of the number of times the data are more likely to be under one of the models (either the null model or the alternative one). The test is also related to Mutual Information, a concept initially introduced by Fano and Hawkins (1961) as a measure of the interdependence of two signals in a message. If we apply this concept to our problem, the measure of Mutual Information between two concepts x and y, $I(x, y)$, compares the joint probability of occurrence of those concepts with the probability of independent occurrences (Church and Hanks, 1990). High values of I indicate a strong relationship between both concepts, while values of I close to

0 indicate independence between them. The application of the G-Test is a useful way to transform the concept of Mutual Information into a statistical test based on the comparison of actual observations to a particular null model. In that case, since the G-Test null model assumes independence of the concepts, a final p-value close to 0 will indicate a statistically significant co-occurrence relationship between the two analysed concepts.

In a similar way to that shown for the Chi-Square test, the definition of the G-Test for n_{12} co-occurrences is:

$$G = 2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right). \tag{3.12}$$

We can substitute the values of O_i and E_i for each of the categories, in a similar way to that in Section 3.3.1.1.

$$G = 2N_{12}\log\left(\frac{N_{12}}{\frac{N_1 N_2}{N}}\right) + 2(N - N_{12})log\left(\frac{N - N_{12}}{N - \frac{N_1 N_2}{N}}\right). \tag{3.13}$$

Finally, by using the proposed values of p_i and x previously defined and simplifying equation (3.13), we obtain

$$G = 2N\left[x\log\left(\frac{x}{p_1 p_2}\right) + (1-x)\log\left(\frac{1-x}{1-p_1 p_2}\right)\right]. \tag{3.14}$$

3.3.1.3 Generation of the P-Value

The last step in the process of building these alternative co-occurrence graphs is the calculation of a p-value which indicates whether two particular concepts, c_1 and c_2 are co-occurring a statistically significant number of times in the corpus (p-value close to 0), and hence a link between them has to be established in the graph. The calculation of p-value given either χ^2 or G, depends on the degrees of freedom d of the particular case of study. In this problem, we already stated that $d = 1$. Given these parameters, the p-value can be calculated as follows:

$$p(X, d) = \left[2^{d/2}\Gamma\left(\frac{d}{2}\right)\right]^{-1} \int_X (t)^{\frac{d}{2}-1} e^{-\frac{t}{2}} dt, \tag{3.15}$$

where X is the result of the test statistic, which in this case can be both χ^2 or G, d is the number of degrees of freedom (in this case $d = 1$) and Γ is the factorial function, generalised to real and complex arguments:

$$\Gamma_x = \int_0^\infty t^{x-1} e^{-t} dt. \tag{3.16}$$

We have used one of the many implementations that can be found online of these formulas for obtaining the p-value from χ^2 or G, and d, for performing this final step of the statistical procedure[i]. After the generation of a p-value for the co-occurrence of every pair of concepts in the corpus, we have all the information needed for building the co-occurrence graph. Weights of the links between nodes in the graph will be calculated in the same way as described in Section 3.2, that is, a threshold p_0 will be defined for a particular graph, and the value of the weight w_{ij} of the edge that connects nodes i and j in the graph will be $w_{ij} = \log(p_0/p_{ij})$, where p_{ij} is the p-value for the co-occurrences of that particular pair of concepts.

3.3.2 Community extraction

Once that we have generated the co-occurrence graph, we are interested in extracting communities of concepts, that is, subgraphs that represent some kind of structural or dynamic affinity. The concepts within the same community will share characteristics that, in our case, may lead to represent specific meanings or senses. In order to generate these communities we have selected the Walktrap algorithm (Pons and Latapy, 2005), based on a random walker that visits the nodes in the graph and tends to get more easily trapped in densely connected subgraphs, which eventually will become the communities. Following this idea, the probability that this random walker goes from node i to node j, or viceversa, in exactly t steps, $P^t_{ij} = P^t_{ji}$ will be high if both nodes belong to the same community. Also, the probability of going from node i to node l or from node j to node l using t steps will be similar ($P^t_{il} \approx P^t_{jl}$) if i and j belong to the same community. With this basis, a measure for the distance between two nodes i and j is presented:

$$d(i,j) = \sqrt{\sum_{l=1}^{n} \frac{(P^t_{il} - P^t_{jl})^2}{k_l}} \tag{3.17}$$

The generation of communities is particularly efficient in this algorithm, given the possibility of generalising equation (3.17) and calculating the distance between already created communities. This way, communities can be merged in order to fulfill the requirements of the final clustering.

Although there exist many more algorithms that can perform this step of community extraction, for the purposes of this chapter we will focus only on Walktrap, since we are not

[i]The implementation is available at http://www.vvlasov.com/2013/06/how-to-calculate-pvalue-from-chisquare.html

interested in analysing the performance of these particular algorithms, but the performance achieved by each of the alternative co-occurrence measures proposed, in comparison with our technique.

3.3.3 Selected Task

The task that we have selected for comparing the statistical methods proposed in this chapter is task 14 of the SemEval 2010 competition: Word Sense Induction and Disambiguation (Manandhar et al., 2010). More particularly, we will focus on the Word Sense Induction part of the task: in a training phase, a set of documents (sentences or paragraphs) containing ambiguous words (nouns or verbs) were generated. For this purpose, for each possible sense of an ambiguous term a query was created, which contained the ambiguous term and a set of unambiguous terms related to that specific sense. This query was used for performing a search through a search engine, and downloading the obtained documents. Hence, it could be assured that the documents retrieved using that query were specifically related to the particular selected sense. A total of 50 nouns and 50 verbs were selected to be the ambiguous words, and a maximum of 1,000 documents were generated for each of them, related to different senses of the ambiguous word. The systems were asked to generate clusters for the different senses induced using this training dataset. In the testing phase, new unseen instances of the same ambiguous nouns and verbs were given to the systems, which had to tag each instance as belonging to one of the clusters (induced senses) generated in the training phase.

The evaluation was conducted following two different measures, both of them based on the comparison between the clustered instances presented by each of the participating systems, and the Gold-Standard containing the correct clustering of the test instances. However, in this chapter we will only use the proposed F-Measure: for its calculation, each cluster C_i proposed by a system was divided into $\binom{|C_i|}{2}$ instance pairs, and each cluster G_i from the Gold-Standard into $\binom{|G_i|}{2}$ instance pairs. Precision was achieved by obtaining the ratio between the common instance pairs and the total number of instance pairs in the Gold-Standard, while recall was represented by the ratio between common instance pairs and the total number of instance pairs in the clusters proposed by the system. The final F-Measure was obtained through the harmonic mean between precision and recall, $F = \frac{2PR}{P+R}$. For more information about the task, see the work by Manandhar et al. (2010).

3.3.4 Method and Results

The documents provided within the training phase represent the source of information that we will use for creating our co-occurrence graphs. However, considering that the main purpose of this chapter is to compare our technique with the alternative methods proposed in Section 3.3.1, we will not consider all the generated documents. Instead, we will select a random subset of 200 documents per ambiguous word, for a total of 20,000 documents, which will be used for generating the co-occurrence graphs. It is important to notice that considering the process of creation of the documents, we can assume that our hypothesis of coherence inside a document is fulfilled in this task, and hence the CO-Graph technique can be applied. Due to the reduction of documents used for creating our graphs, the results obtained by our techniques are likely to be far from those obtained by the systems participating in the task. However, they will represent a good evaluation setting for comparing the proposed techniques. In a pre-processing phase, we perform a Part-Of-Speech tagging of the proposed documents through the TreeTagger tool (Schmid, 1994), and the co-occurrence graphs are created following the proposed algorithms, and using only nouns and verbs as representative words. Finally, communities are extracted from the co-occurrence graphs using the Walktrap algorithm. Each generated community represents a particular sense that can be found in the corpus.

For assigning a specific cluster for each new instance of the testing set, the community with highest overlap value is selected, that is, the community which present a higher number of words from the test instance. As the number of communities created may be very high (some communities contain only a few words and others present hundreds of words), a post-processing step is carried out, by first selecting the minimum percentage C of instances (with respect to the total number of test instances, I) assigned to a particular community for it to be maintained. That is, after the assignation process, every community that does not present at least $(\frac{C}{100})I$ instances is removed, and each of its instances reassigned to the community presenting the second highest overlap value for it. The only case in which a community that presents less than $(\frac{C}{100})I$ instances is maintained is when its instances cannot be reassigned (they do not overlap with any other community). Then, the post-processing stops when each community is related to at least that minimum value of instances, or to instances that cannot be reassigned.

In the first experiment we want to analyse the behaviour of the proposed technique, CO-Graph, as we vary the threshold used for determining the restrictiveness of the graph (p_0). Starting with a maximum $p_0 = 10^{-2}$ as stated in Section 3.2, as we decrease this value the graph becomes more restrictive (it is more difficult that the number of co-occurrences of a pair of concepts is statistically significant). We will also vary the minimum percentage of the total number of instances that must be assigned to each final community at the end of the post-processing step. The results of this first experiment are shown in Table 3.1.

C(%)	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
5	0.468	0.481	0.446	0.477	0.469
6	0.469	0.482	0.456	0.479	0.475
7	0.471	0.484	0.461	0.482	0.479
8	0.473	0.486	0.463	0.487	0.484
9	0.474	0.488	0.475	0.488	0.487
10	0.478	0.490	0.478	0.489	**0.492**

Table 3.1: Results (F-Measure) for the WSI task 14 of SemEval 2010, obtained using the CO-Graph technique for building the co-occurrence graph, as we vary the threshold p_0 (first row) and the final number of instances per community (first column), expressed in % of the initial number of instances.

As we can observe, results improve for all the studied thresholds as we increase the final number of instances per community (that is, the number of final communities representing word senses decreases). Hence, we need to perform a more exhaustive analysis on this parameter for comparing the different techniques described in this chapter. Regarding the threshold selected for the comparison, although results are better for most of the cases when $p_0 = 10^{-3}$ and $p_0 = 10^{-5}$, for the maximum value of C tested in this experiment ($C = 10\%$) we observe that the threshold that offers the best results is $p_0 = 10^{-6}$. Hence, we will select this last threshold value for the experiment that performs the comparison between the three co-occurrence techniques.

Figure 3.1 shows the results of the developed technique CO-Graph in comparison to the alternative methods described in Section 3.3.1 (Chi-Square and G-Test), as we vary the final number of instances per community, with a fixed threshold value of $p_0 = 10^{-6}$.

The results clearly show that the CO-Graph technique is able to achieve better results than both the G-Test and the Chi-Square statistical test of independence for the considered task. F-Measure increases for all the analysed techniques as we increase the final number of instances per community (and hence the number of possible senses per ambiguous word decreases), until we reach $C = 45\%$, that is, each community should have almost half of the instances. The results start decreasing at the same point for all the considered techniques, although the decreasing is faster in the Chi-Squared test.

3.4 Conclusions

In this chapter we have presented the mathematical background behind the construction of the co-occurrence graph, as well as a comparison between the statistical model on which the calculation of the significance of co-occurrences between concepts is based, and other statistical tests of independence also offering values for this statistical significance. This

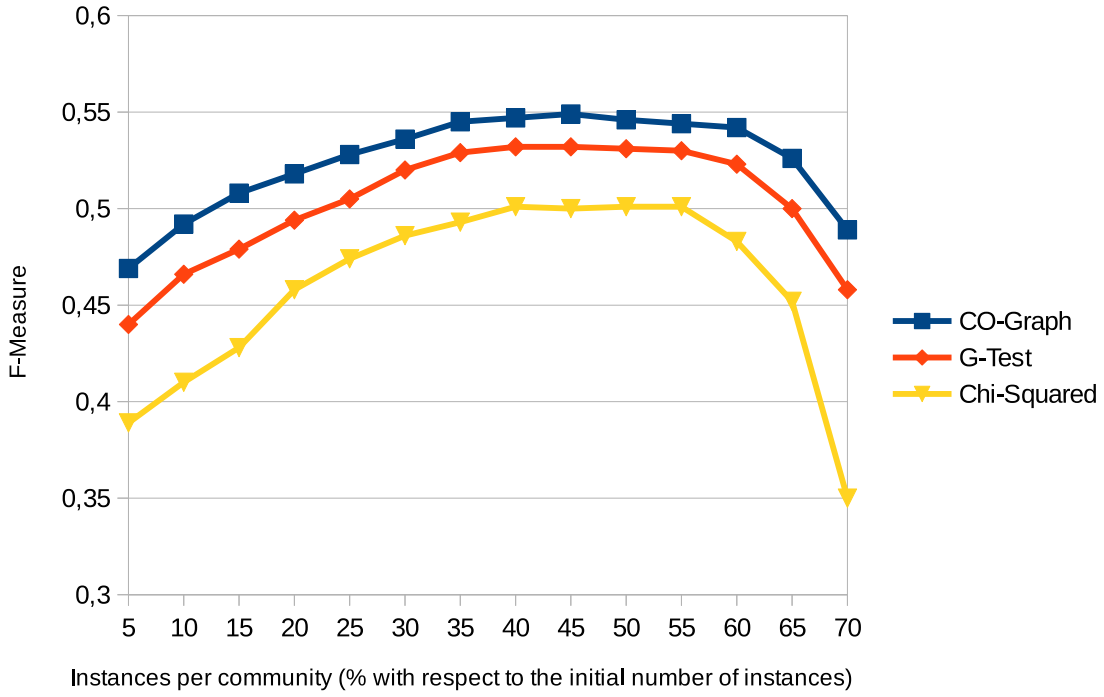


Figure 3.1: Comparison of F-Measure for the proposed technique (**CO-Graph**), the **G-Test** and the **Chi-Square** statistical test of independence. Evolution of the F-Measure as the final number of instances per community (% with respect to the initial number of instances) increases.

comparison has been carried out through the application of the proposed techniques to a Word Sense Induction task. The obtained results validate the proposed CO-Graph technique in terms of the algorithms presented for calculating this co-occurrence significance. Although offering close results, the alternative statistical methods explored in the experiments are not able to achieve the same results as the CO-Graph technique. For the sake of understanding, some ideas such as the creation of communities and the restrictiveness of the graph as we vary the threshold p_0 have been presented, although they are not fully related to the purposes of this chapter. We will explore those ideas in more detail in subsequent chapters of this thesis.

4

CROSS-LINGUAL WORD SENSE DISAMBIGUATION

If you talk to a man in a language he understands, that goes to his head. If you talk to him in his own language, that goes to his heart.

Nelson Mandela

Contents

4.1 Introduction

Cross-Lingual Word Sense Disambiguation (CLWSD) aims to determine the most suitable translation for a given word from a source language to a target one. CLWSD tries to deal with some of the difficulties of WSD, such as the scarcity of sense inventories and sense tagged corpora, by taking advantage of the shared meaning between parallel texts. Parallel corpora are considered the source of knowledge to perform the disambiguation in this work. These corpora are good resources not only for performing CLWSD, but for NLP in general (Resnik, 2004), since parallel translations share hidden meaning that can be useful for extracting knowledge about a language, from another language richer in resources.

In this chapter, we will explore the use of co-occurrence graphs, built through the technique described in Chapter 3, for solving CLWSD tasks. The robustness of the technique will be tested by applying it to many different languages. For the purposes of this particular task, our co-occurrence window will be the whole document, what helps to improve the coverage of the system, while as we stated in previous chapters, the precision is maintained by excluding those words whose co-occurrence rate is below the threshold defined by a null model of distribution of concepts among the set of documents.

As we briefly introduced in Chapter 2, in this kind of tasks the disambiguation process usually relies on a bilingual dictionary that proposes a set of potential translations for an ambiguous word, which represent the ambiguity to be solved. Any system attempting to perform CLWSD has to choose, for each instance of an ambiguous term, which translation is the most suitable, among those provided by the dictionary. This dictionary will be built in an automatic, unsupervised way, in order to maintain the full unsupervised nature of the proposed technique. Considering the proposed resources (co-occurrence graph and bilingual dictionary), there exist many algorithms and techniques that use the information offered by them for performing the final disambiguation. Some of those algorithms will be analysed, and the final results of the system will be compared to baselines and other unsupervised systems that deal with the same problem, for proving the effectiveness and robustness of our system.

The rest of the chapter is organized as follows: Section 4.2 defines the problem to be solved. Sections 4.3 to 4.6 describe the proposed system, detailing the different steps involved in the disambiguation process. Section 4.7 shows the evaluation framework and criteria, and illustrates the different experimental results obtained by the system and their comparison to other systems. Finally, conclusions are gathered in Section 4.8.

4.2 Problem Definition

Our main objective is to find the most suitable translations for a given word in a provided context, from a source language to a target one. The context is represented by a sentence in which the target word can be found.

For example, in the sentence "*But **coach** experts say it hasn't been proved that belts are safer*", with English as source language, the word "***coach***" is the target word, while the rest of the words in the sentence represent the context. Taking, for instance, German as target language, the target word should be translated as "*Bus*", "*Reisebus*", "*Omnibus*", "*Linienbus*" or "*Busunternehmen*". CLWSD aims to select the most appropriate of those translations.

This problem has been addressed in two different editions of the SemEval competition: task 3 of SemEval 2010 (Lefever and Hoste, 2010a) and task 10 of SemEval 2013 (Lefever and Hoste, 2013). Hence, we will use the test datasets proposed in those competitions for evaluating our system. In the CLWSD problem, some source of information is used to extract the knowledge about the domain. The knowledge base in this work is represented by the Europarl corpus (Koehn, 2005), which will be described later.

4.3 System Description

The system presented in this chapter consists of two main modules, each of them representing a phase in the disambiguation process: the first module transforms the base of knowledge, represented by a parallel corpus, in data structures that can be used by the disambiguation system. The second step of the process is the disambiguation itself. In the first module, we automatically extract bilingual dictionaries from the base corpus. Our system selects the most suitable translations among those provided by the dictionary for a target word in its context. The selection is done according to the information which arises from a co-occurrence graph. In this section, we describe the approach followed to construct the bilingual dictionaries and the co-occurrence graph, as well as the different techniques that have been tested to identify sets of highly related nodes in the graph, which can be viewed as words related to a particular sense.

Figures 4.1 and 4.2 illustrate the complete system. Figure 4.1 shows the construction of the two data structures used in our system: the bilingual dictionary and the co-occurrence graph. The bilingual dictionary is obtained from the original parallel corpus, both in the source language (always English) and in the target language (Spanish, French, Italian, German or Dutch). The corpus written in the target language is also taken as base for building the co-occurrence graph, after a pre-processing step and an optional filtering process. In

Figure 4.2 we can observe how, given a test sentence, all the algorithms that we have explored for performing the final step in the disambiguation process, make use of the bilingual dictionary and the co-occurrence graph for providing an initial set of translations for the target word. This set is then combined with the translation probabilities extracted directly from the bilingual dictionary for generating the final output of the task. Each of the disambiguation algorithms, as well as the process of combining the initial output with the translation probabilities, will be explained later on.

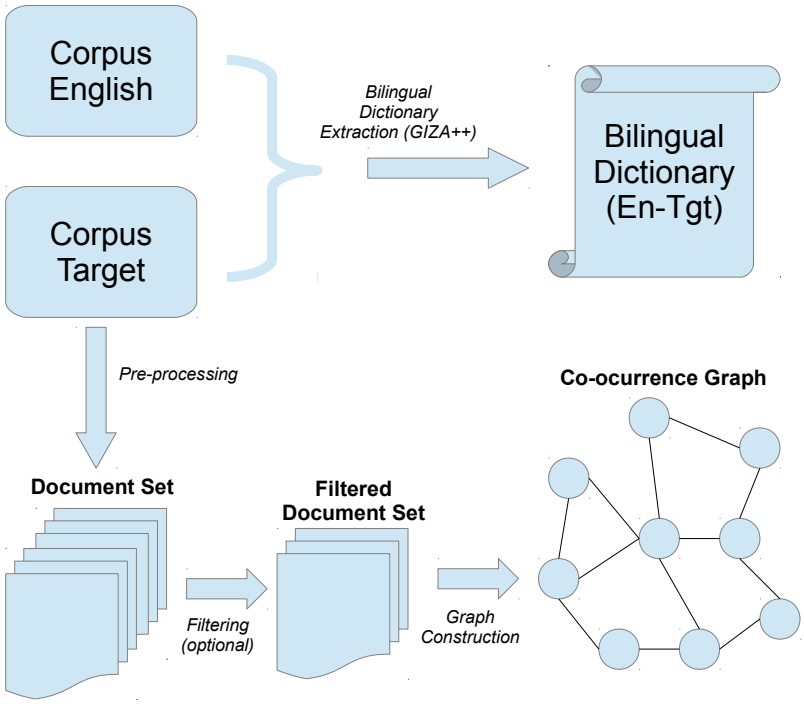


Figure 4.1: Construction of the bilingual dictionary and the co-occurrence graph.

4.4 Bilingual Dictionary Extraction

We need to obtain bilingual dictionaries for all the proposed target languages, with English as source language, in an automatic way. For this purpose, we use GIZA++ (Och and Ney, 2003), an automatic tool that allows us to align at word level two parallel corpus, originally aligned at sentence level. The input of the GIZA++ tool consists of two files, each one of them representing a whole corpus written in a specific language. In these files raw text must be introduced, represented by one word per line, without XML tags or any other marks apart from the text to be aligned. We apply this tool over the sentence-aligned Europarl corpus of a pair of languages, and use the intersection of both translation directions. This way, we obtain bilingual dictionaries from English to German (En-De), from English to Spanish (En-Es), from English to French (En-Fr), from English to Italian (En-It), and from English to Dutch

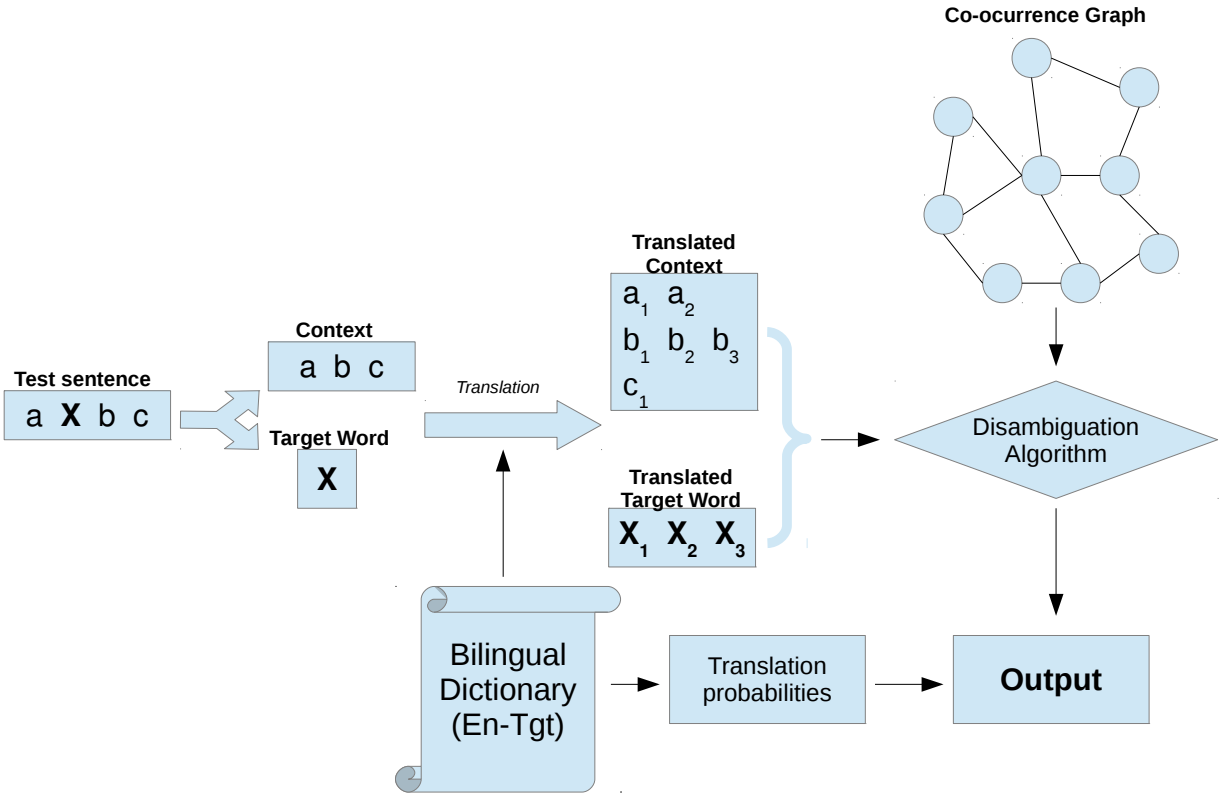


Figure 4.2: Disambiguation process. Each of the analysed disambiguation algorithms makes use of the bilingual dictionary and the co-occurrence graph as sources of information for disambiguating each target word in each test sentence. "Tgt" stands for "target".

(En-Nl). It is important to remark that the GIZA++ tool provides information about the probability of occurrence of each translation of a target word. These translation probabilities, as we briefly explained in Section 4.3 and Figure 4.2, will be used as a prior probability for the translations provided by the bilingual dictionary. Specifically, this information given by GIZA++ will be combined with the initial output of our system, called CO-Graph, for generating the final output for every test instance.

Table 4.1 and 4.2 show some statistics extracted from the bilingual dictionaries obtained with the GIZA++ tool. Table 4.1 contains the number of entries of each dictionary, the maximum number of translations of a single entry, as well as the average number of translations per entry that can be found. Table 4.2 shows the number of translations provided by each of the bilingual dictionaries, for each target word in the test dataset of the SemEval 2010 and SemEval 2013 competitions.

As we can observe in the tables, although the average number of translations per word is low, there exist words in the dictionaries that present a very large number of potential translations. In general, most of the target words in the test dataset present many translations, due to the automatic nature of the GIZA++ tool. We can then prune the dictionary and only consider

	En-De	En-Es	En-Fr	En-It	En-Nl
Number of Entries	32,498	34,815	34,029	34,152	33,751
Max number of translations	946	1,344	1,347	1,534	1,290
Avg number of translations	8.46	7.51	7.23	8.03	9.79

Table 4.1: Statistics from the bilingual dictionaries obtained through the GIZA++ tool, from English to German (column En-De), Spanish (column En-Es), French (column En-Fr), Italian (column En-It) and Dutch (column En-Nl).

Word	En-De	En-Es	En-Fr	En-It	En-Nl
coach	33	8	13	25	30
education	135	52	35	51	137
execution	40	30	21	33	50
figure	158	146	138	143	186
job	161	133	143	132	208
letter	76	46	54	62	67
match	82	101	96	100	91
mission	99	35	35	40	116
mood	20	32	22	28	34
paper	78	64	64	63	88
post	95	72	70	81	95
pot	13	21	13	16	15
range	103	100	99	111	109
rest	51	87	74	94	92
ring	38	34	35	40	41
scene	57	46	43	56	67
side	160	191	221	206	205
soil	26	10	13	16	35
strain	31	48	44	44	47
test	112	89	71	91	130

Table 4.2: Number of translations of the words in the test dataset, for each bilingual dictionary obtained through the GIZA++ tool.

those translations with highest probability. Some tests have shown that a pruning value of ten translations per word provides the best results. Hence, our system will have to select the most suitable translations for a given word in a given context, among a set of ten translations.

Figure 4.3 illustrates a specific example of the disambiguation process. In this case, we want to disambiguate the word "coach" in a specific sentence, from English to Spanish. As we can observe, words surrounding the target words are considered as context. The bilingual dictionary provides the possible translations of the target word, as well as the translations of the words in the context. This information, together with the knowledge embedded in the co-occurrence graph, will allow us to perform the disambiguation, using one of the algorithms that will be explained later on.

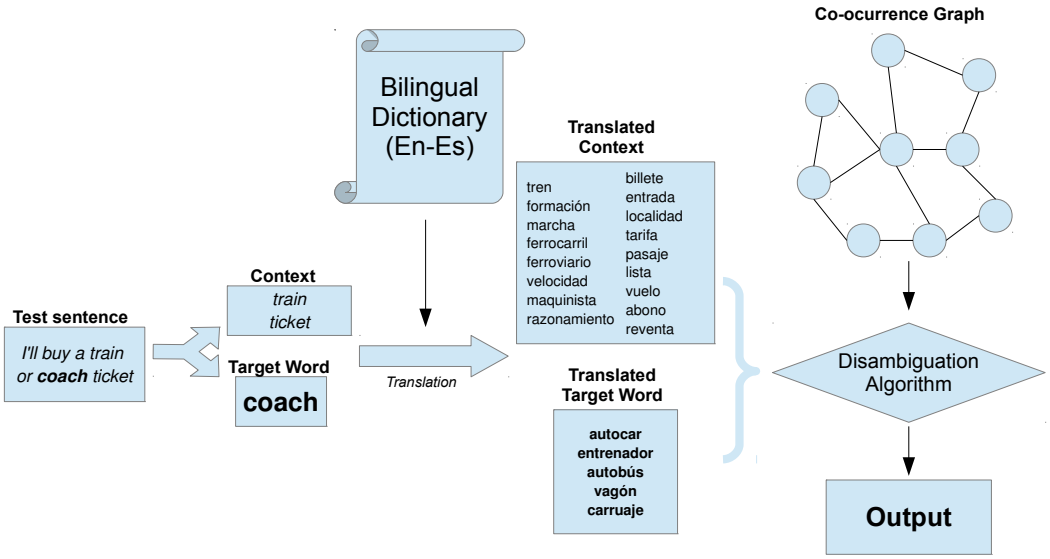


Figure 4.3: Example of the disambiguation process of a sentence containing the target word "coach", with Spanish as target language.

4.5 Knowledge Representation

4.5.1 Corpus Pre-processing

Although the Europarl parallel multilingual corpus, extracted from the proceedings of the European Parliament and taken as knowledge base for the task, is presented in many languages, only those proposed in the evaluation tasks are taken into account, namely English, Spanish, French, Italian, German and Dutch.

We split the initial corpus, divided in XML-tagged documents, by detecting the interventions of different members of the Parliament. Each intervention, labelled with the "speaker" tag, will become a document to be used later on by our algorithm. In this way, we intend to fulfil our hypothesis that the words appearing in the same document are likely to be related to the general sense of the document.

The words inside the documents need to be lemmatized and tagged according to their Part-Of-Speech (POS) tag. The lemmatization and POS tagging is automatically performed through the use of the TreeTagger tool (Schmid, 1994).

4.5.2 Document Filtering

We have considered two different ways of building the co-occurrence graph from the documents extracted from the original corpus. The first approach takes into account all the documents for building the graph, hence the same graph will be used for disambiguating any word from the test dataset. This approach will be denoted as "Complete graph approach" along the rest of the chapter. The second approach is based on the belief that more specific graphs will provide better results when performing the disambiguation. For this purpose, we build a specific graph for each of the target words appearing in the test dataset, by removing, from the original document set, those documents that do not contain any of the possible translations of the target word. Hence, we will obtain as many graphs as target words exist in the test dataset, and we will use, for each sentence, the graph that corresponds to the target word, for disambiguating it. This second approach will be denoted as "Word-based graph approach" along the rest of the chapter.

4.5.3 Co-occurrence Graph Construction

From the tagged documents in the target language, we are now able to build the co-occurrence graph that will allow us to perform the disambiguation. As we consider in our initial hypothesis, the appearance of a word in a document is likely to be related to the general sense of the document (intervention of a member), but not necessarily. Regarding the purposes of this task, the concepts that will populate our co-occurrence graph will be words, and more particularly nouns from the text. The use of words belonging to a different part of speech, such as adjectives or verbs, may lead to the construction of unmanageable graphs in terms of size (number of nodes and links between nodes). Some experiments have been conducted for testing this issue, although the results do not improve those obtained using only nouns as nodes of the graph. These experiments will be shown in further chapters of this thesis. Considering this, we extract from the tagged documents all the words marked as "noun" by the POS tagger. This way, each of the documents considered for building the graph will be represented by a set of nouns. Once that we have those elements, we are able to build the co-occurrence graph following the technique described in Chapter 3.

4.6 Target Word Disambiguation

The construction of the co-occurrence graph gives us a structured representation of the knowledge inside the corpus. We now need to select from the graph those nodes closely related that can be considered to be related to the same sense. Although there exist many possible implementations of this step, in this work we will study three different techniques for determining the most suitable translations for a given word in a given context: Community detection, PageRank algorithm and Dijkstra's algorithm.

4.6.1 Community Detection and Community Graph

A community is a sub-graph whose nodes present some kind of structural or dynamic affinity. In this technique, we assume that words belonging to the same community share a common sense, different from those represented by other communities. There exist many different community extraction algorithms. In this work, we use two algorithms widely explored in the literature, and compare their results:

- **Walktrap**: The Walktrap algorithm (Pons and Latapy, 2005), already presented in Section 3.3.2 of Chapter 3, is based on the fact that a random walker that jumps between nodes inside the graph, will get more easily trapped in those sub-graphs that are densely connected. These sub-graphs would then become the communities. It is a particularly efficient community-based algorithm, since it can be easily generalized to a coarse-grained structure, and hence communities can be merged into bigger modules for allowing a faster computation.
- **Chinese Whispers**: The Chinese Whispers algorithm (Biemann, 2006) is a simple yet efficient technique that assigns each vertex to a community in a bottom-up fashion. In the first step, the algorithm assigns a distinct class to each vertex. Then, the nodes are iteratively assigned to the class that contains the strongest neighbours of the analysed node (those with highest weights in edges linked to the current node).

Figure 4.4 shows an example of the differences between communities when using both algorithms. We have selected the word "entrenador" in Spanish, which is a translation of the word "coach", referred to a person that trains an athlete or team. As we hypothesized, words in the communities tend to be related to this particular sense of the word "coach", such as "futbol" ("football" or "soccer"), "golf", "rugby", "entrenamiento" ("training session"), "arbitro" ("referee"), "jugador" ("player"), "estadio" ("stadium"), "campeonato" ("championship"). As we can observe, both algorithms generate a similar community, although the one provided by the Chinese Whispers algorithm is smaller.

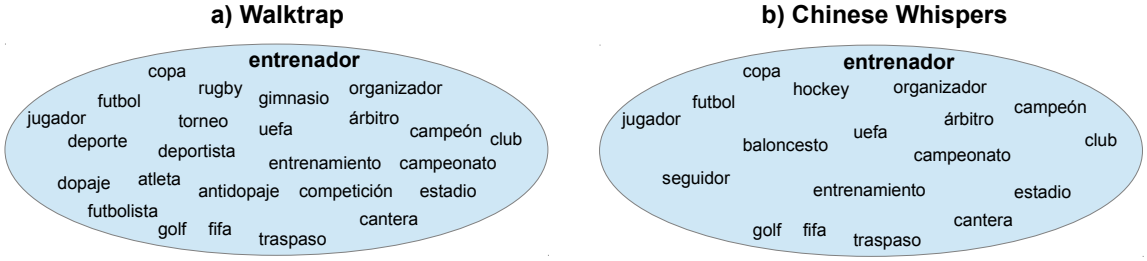


Figure 4.4: Communities containing the word "entrenador" as translation of "coach" in Spanish: (a) Walktrap algorithm; (b) Chinese Whispers algorithm.

With the communities obtained by the algorithm, we build a new graph, called community graph (CG). In this graph, each community is represented by a node, and an edge will be added linking communities (nodes) C_1 and C_2 if and only if any word $x \in C_1$ is linked in the co-occurrence graph to any word $y \in C_2$.

Context surrounding the target word is the only additional information that can be used to perform the disambiguation. As we stated in Section 4.5.3, the co-occurrence graph has been built using only nouns, thus all the remaining words (adjectives, verbs, ...) can be removed from the context.

The next step is to identify, inside the community graph CG, those communities that contain at least one of the translations, either from words of the context or from the target word. As a

result, we obtain two sets of communities: set M_T includes communities that contain at least one translation from the target word, and set M_C is composed of communities containing at least one translation from any word of the context. Through the community graph we can calculate the distances between any community $M_C^i \in M_C$ and any community $M_T^j \in M_T$. Since a translation of a target word can belong to the same community that a translation of a context word ($M_C^i = M_T^j$), the distance in that case would be 1, which is the minimum distance we consider. In any other case, we will add the number of links in the shortest path between M_C^i and M_T^j. Hence, if the path between M_C^i and M_T^j contains one link, the distance between them, for our purposes, would be 2, if the path contains 2 links, the distance would be 3, and so on.

Our hypothesis for this algorithm is that the translation of the target word that is nearer (in average) to the translations of the context words, is more likely to be the most suitable translation for that target word in that context. Hence, we establish a formula for ranking the potential translations of the target word, based on two factors: the score of a translation is inversely proportional to the distance between the community to which it belongs and any community containing context translations, in order to give greater emphasis to first-order co-occurrences (Schütze, 1998), but directly proportional to the number of context translations inside the community. Thus, the weight or score of a translation of the target word, w_t, will be given by:

$$w_t = \max_{M_C^i \in M_C} \frac{A_C^i}{(d_{M_C^i M_T^t} + 1)} \tag{4.1}$$

where A_C^i is the number of context translations inside M_C^i, and $d_{M_C^i M_T^t}$ is the distance (number of steps) between M_C^i and M_T^t, that is, the community in which translation t is located. By ranking the scores of all the possible translations for the target word given by the dictionary, the system can propose the most suitable ones as a solution.

Figure 4.5 illustrates the algorithm based on communities and contains an example of its behaviour, as explained above. In the example, links between nodes of the community graph do not necessarily represent paths containing one only link, but any possible value of $d_{i,j}$. Hence, in that example word X_1 will have a weight $w_1 = \max(\frac{3}{d_{1,1}+1}, \frac{1}{d_{1,2}+1}, \frac{2}{d_{1,3}+1})$. Word X_2 will have the same weight, $w_2 = w_1$, since X_1 and X_2 belong to the same community. Finally, word X_3 will have a weight $w_1 = \max(\frac{3}{d_{2,1}+1}, \frac{1}{d_{2,2}+1}, \frac{2}{d_{2,3}+1})$.

4.6.2 PageRank Algorithm

The PageRank algorithm (Brin and Page, 1998) is used over a graph for ranking the importance of each of its nodes. This algorithm has been widely used in the last years for performing Word Sense Disambiguation (Agirre and Soroa, 2009; Mihalcea, 2005; Navigli

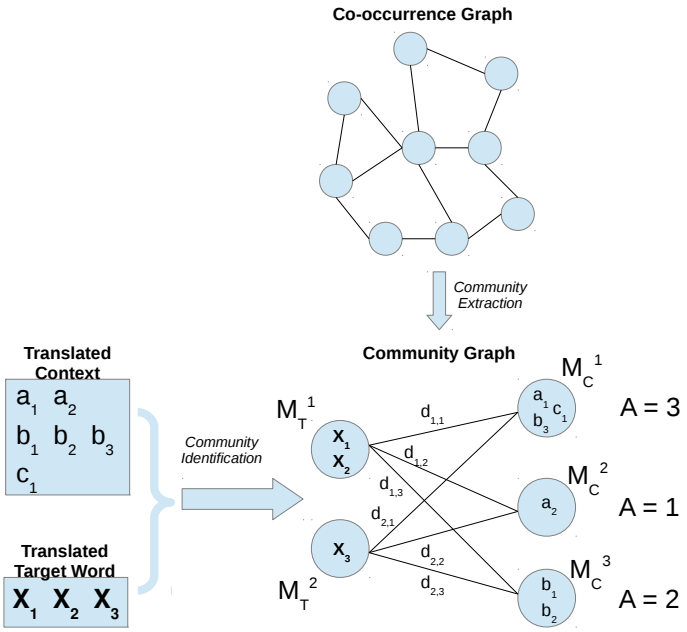


Figure 4.5: Diagram and example of the community-based algorithm. The community graph is extracted from the co-occurrence graph, and used to compute the distances between words from the context and the target word.

and Lapata, 2010). The PageRank calculation for the whole graph can be performed through the following formula:

$$P = dMP + (1 - d)v \tag{4.2}$$

P is a vector with the PageRank values for each node, d is a constant called "damping factor" and usually set to 0.85, M is the matrix representing the out-degrees of the nodes, and v is a $N \times 1$ stochastic vector, being N the number of nodes in the graph. By means of v, the probability of randomly jumping into a node of the graph can be distributed among the nodes of the graph in different ways. In this work, we will explore two approaches for applying the PageRank algorithm:

- **Basic PageRank**: All the members of vector v will have the same value, $v_i = \frac{1}{N}$. Therefore, in this approach the context of the sentence in which the target word appears is not taken into account.
- **PageRank with Priors**: In this case, the vector v will be used for giving more importance to those words surrounding the target word in a specific context, in a similar way to that explained by Agirre et al. (2013). If there are C words in the translated context of a specific sentence, the values of members of vector v will be $v_i = \frac{1}{C}$ if node i represents the translation of a word of the context, and 0 otherwise.

Once that we have calculated the PageRank (either basic or with Priors) of our co-occurrence graph, we can determine the most suitable translations of a target word by simply selecting those translations with higher values of PageRank.

Figure 4.6 shows the behaviour of the PageRank algorithm for disambiguation. The translations of the target word are sought inside the graph, and their weights will correspond to those assigned by PageRank, $w_1 = w_{1\_PR}$, $w_2 = w_{2\_PR}$ and $w_3 = w_{3\_PR}$.

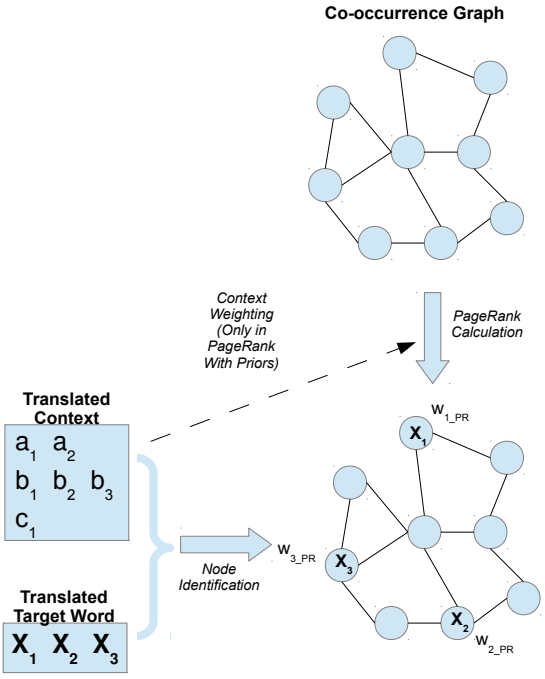


Figure 4.6: Diagram and example of the PageRank algorithm. The translation of the context is used only if we are performing the "PageRank with Priors" approach.

4.6.3 Dijkstra's Algorithm

The shortest path from node i to node j of a graph can be calculated through Dijkstra's algorithm (Dijkstra, 1959), which uses weights of the links for selecting a path. Through this algorithm, we can calculate the shortest distance between the translation of a word of the context, and a translation of a target word, and use this information for ranking those translations. Since weights in the graph represent the importance of a link between two words, we will assign, for each link, the inverse of its original weight for obtaining the minimum distances.

For assigning a value to the influence that a context word has in the selection of a particular translation of the target word, we retrieve the original weights (representing the importance of a link between two words) of the links in the graph and sum all the values of the edges

involved in the shortest path. Then, this final sum is divided by the number of edges in the path. Hence, for each translation of each context word, t_c, and each translation of the target word, t_w, we obtain a value related to the shortest path between those words in the original graph. Then, the score of each possible translation of the target word will be the highest weight that is assigned in this step, this is, the highest influence given by a context word. By ranking these values, we can determine the most suitable translations for the target word given a specific context.

Figure 4.7 shows the different steps of the technique based on Dijkstra's algorithm, and an example of its behaviour. Dijkstra's algorithm is applied to the co-occurrence graph. Following the above description of the algorithm, if we represent the influence of a translation of a context word, t_c, over a translation of the target word, t_w, in terms of a function $I(t_c, t_w)$, we obtain that $I(a_1, X_1) = \frac{e_{10}+e_{13}}{2}$, $I(a_1, X_2) = e_5$ and $I(a_1, X_3) = e_4$. Hence, considering the translations of context words shown in the example of the figure, the weight of any translation of the target word X_n will be:

$$w_n = \max\left(I(a_1, X_n), I(a_2, X_n), I(b_1, X_n), I(b_2, X_n), I(b_3, X_n), I(c_1, X_n)\right).$$

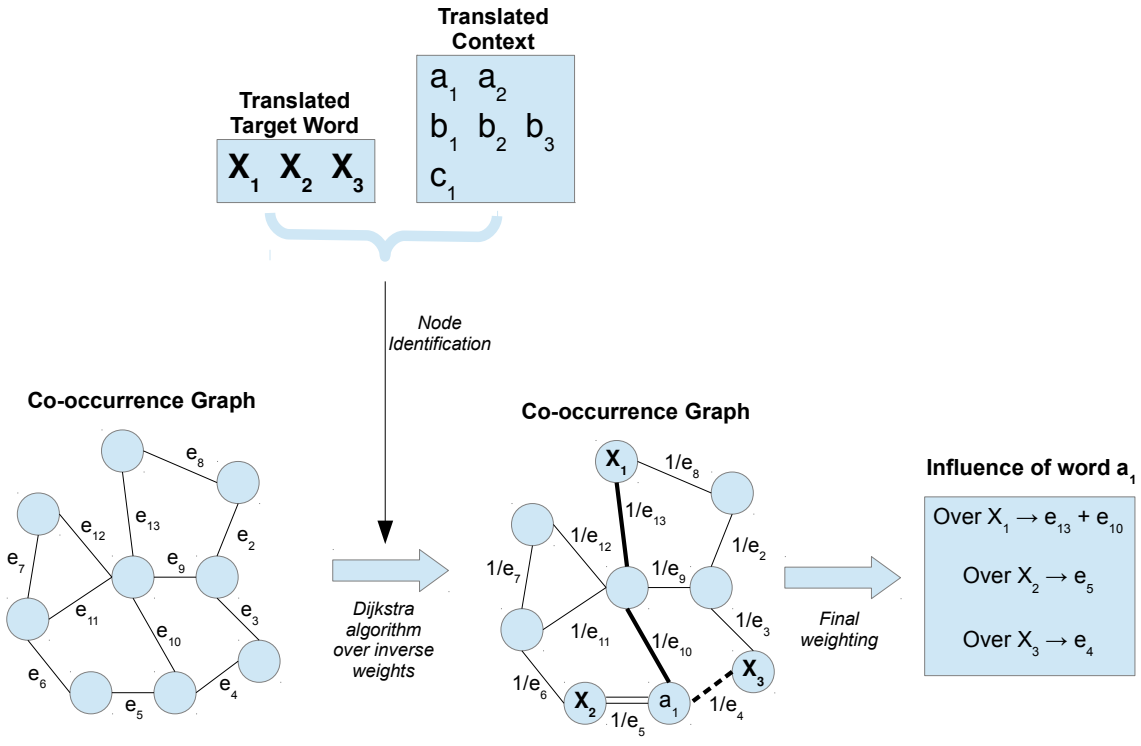


Figure 4.7: Diagram and example of the technique based on Dijkstra's algorithm. The weights of the edges in the co-occurrence graph are inverted for computing the shortest path algorithm. The different lines in the edges after applying Dijkstra's algorithm represent the shortest paths from a_1 to X_1 (continuous line), X_2 (double line) and X_3 (dashed line) respectively.

4.6.4 Output of the System

GIZA++ provides very valuable information about the translations, and it has a different nature than the weights our system assigns to each potential translation. This information is also obtained in an automatic, unsupervised way. This fact suggests that combining the weights obtained by our system, through any of the disambiguation algorithms, with the most probable translations of a target word, should offer good improvements to the final disambiguation. Accordingly, we will assign a final score to each of the ten potential translations provided by the dictionary. This final score will be extracted by multiplying the score obtained by the CO-Graph system, and the probability of translation given by GIZA++. This is, we consider $T = (t_1, t_2, ..., t_n)$ to be the complete set of potential translations provided by GIZA++ for a given target word, where $n <= 10$. Each translation t_i has an associated probability p_i. After applying the disambiguation process, the CO-Graph system assigns a weight w_i to each of the potential translations. The final score of each translation s_i will be given by $s_i = p_i w_i$. In the following experiments we will illustrate the results obtained by applying this combination of the weights of the system and the probabilities of translation. Besides, we will also show results obtained by only using the CO-Graph system, without using this prior probability or back-off given by GIZA++.

4.7 Evaluation

4.7.1 Evaluation Criteria

The evaluation setting followed in our experiments is based on the one proposed in task 3 of SemEval 2010 and task 10 of SemEval 2013 competitions. The disambiguation is performed taking English as the source language, and five different languages as target languages, namely Spanish, French, Italian, Dutch and German. Systems participating in the tasks are asked to propose the most suitable translations for each test sentence in as many languages as possible. Evaluation is carried out, in both tasks, over a test dataset with 20 different words and 50 sentences for each of them, and results are compared against a manually built Gold-Standard containing the most suitable translations for each target word in each sentence.

The Gold-Standard is built from the Europarl corpus. For this purpose, a word-level alignment was performed and manually evaluated for all the sentences of the corpus containing target words, for every pair of languages containing English as source language. After that,

a manual clustering by meaning was carried out, for every target word. The output of this process was a sense inventory, used for annotating the datasets for the tasks (Lefever and Hoste, 2010b).

Annotators of the Gold-Standard used the clustered sense inventory for selecting the most appropriate translations of each target word. The translations are weighted depending on how many annotators selected each of them. Example 4.3 shows the Gold-Standard provided by the annotators for a given sentence in which we can find the target word "***coach***".

(4.3) SENTENCE 2: *A branch line train took us to Aubagne where a* ***coach*** *picked us up for the journey up to the camp.*

coach.n.nl 2 :: bus 3;autobus 3;toerbus 1;touringcar 1;

coach.n.fr 2 :: bus 3;autobus 3;car 3;

coach.n.de 2 :: Bus 3;Omnibus 2;Reisebus 2;Linienbus 1;Reisenbus 1;

coach.n.it 2 :: autobus 3;pullman 2;corriera 2;autocarro 1;pulmino 1;

coach.n.es 2 :: autocar 3;autobus 3;diligencia 1;

Two different evaluation schemes are proposed:

- **Best Evaluation:** The first evaluation scheme asks the systems to propose any number of translations for each target word in each context, but the final score is divided by the number of translations. Hence, the scoring process penalizes the systems proposing too many translations. In our case, we consider only those translations (up to two), among the ten potential translations provided by the bilingual dictionary, whose normalized weight (between 0 and 1), is higher than 0.3 (30% of the total weight). If there are no translations that fulfil this condition, only the translation with the highest weight is proposed, except if there exists a tie. In that case, the two translations with the highest weights are considered.
- **Out-Of-Five Evaluation:** This "more relaxed" scheme expects an output of up to five different translations for each target word in each context, without penalizing the system according to the number of translations. Hence, given that the dictionary constrains the number of translations of each word to ten, our system will have to select, in any of the proposed disambiguation algorithms, five of those potential translations as a solution for each test instance.

The evaluation measure considered in these tasks is F-Measure. Due to the nature of the guessings proposed by our system, the values of precision and recall (and hence the value of F-Measure) are always the same. Then, we will refer only to the F-Measure value for illustrating the results achieved by our system, and for comparing them with other systems participating in the SemEval competitions.

4.7.2 Example of Disambiguation

In this section we present two examples of the disambiguation process for a given word in a given context (test sentence). Since the creation of the co-occurrence graph can not be illustrated in an example due to its size, only the disambiguation step is shown, after the generation of the co-occurrence graph. The technique illustrated in these examples is the community-based algorithm, and the language is Spanish. Both evaluation schemes are presented and the results offered by the CO-Graph system can be compared to the Gold-Standard for the same test sentences. Figures 4.8 and 4.9 show the disambiguation in Spanish of the target word "***test***" and "***strain***", respectively, each of them within a test sentence.

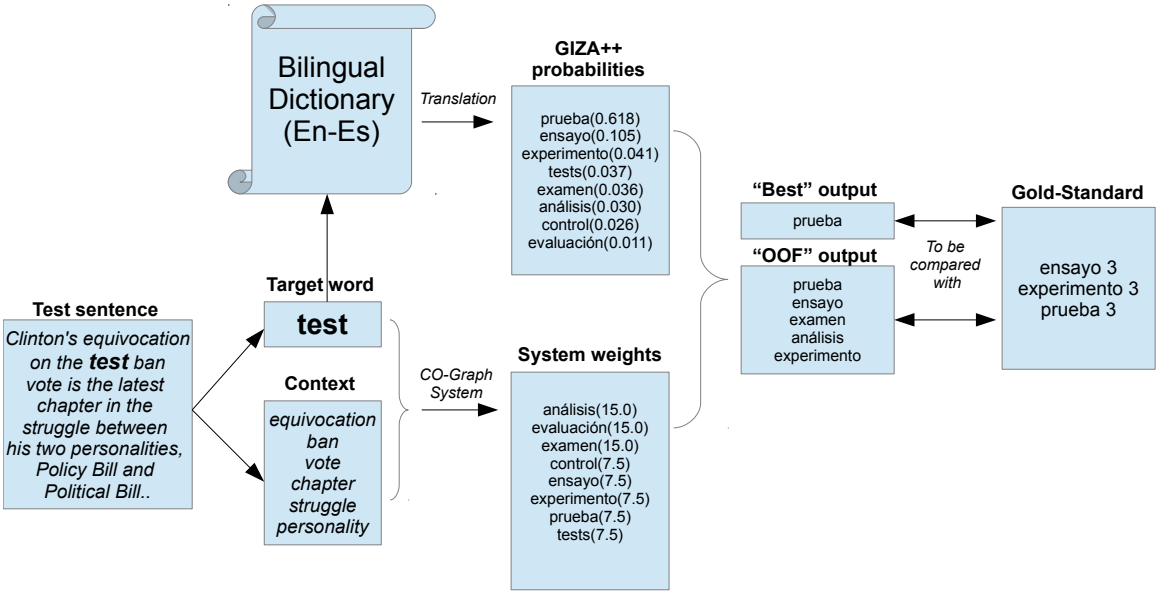


Figure 4.8: Example of disambiguation. The target word is "***test***" and the final language is Spanish.

The target word and the context are separated from the original sentence, and introduced as an input to the CO-Graph system. The system uses the bilingual dictionary and the co-occurrence graph, as well as the community-based technique, for extracting the weights for disambiguation. At the same time, the translation probabilities are obtained from the bilingual dictionary. The final scores are obtained and ranked according to the scoring process stated in Section 4.6.4. Hence, we obtain the final translations for the "Best" and "Out-Of-Five" evaluation schemes. This will be compared against the Gold-Standard for evaluating the disambiguation.

In the first example, we observe that the CO-Graph system assigns two different values (15.0 and 7.5) to the possible translations before using the translation probabilities given by GIZA++. According to those values, the words in the Gold-Standard are not among those

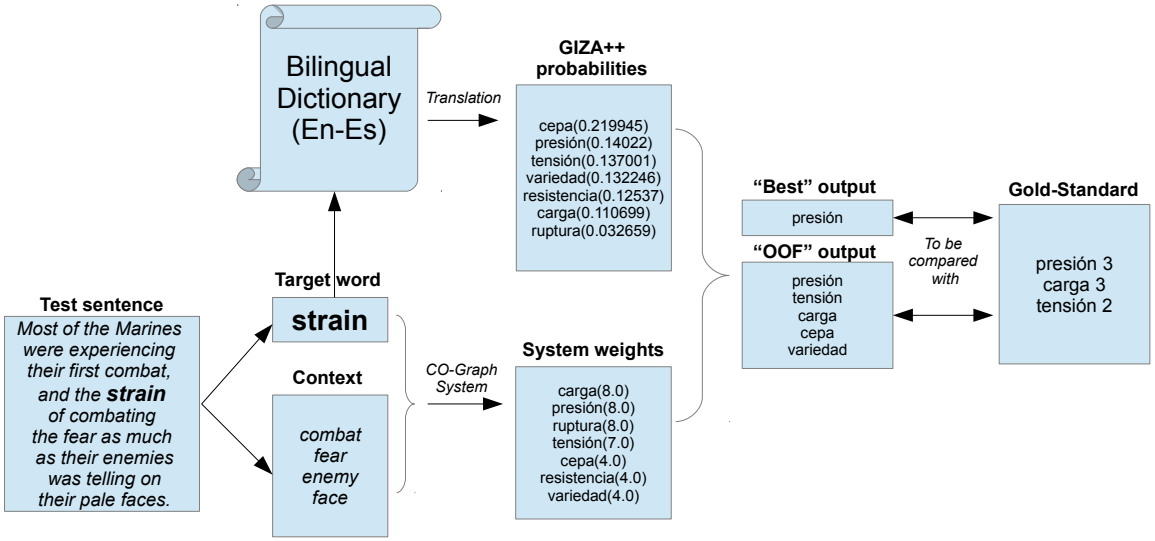


Figure 4.9: Example of disambiguation. The target word is "***strain***" and the final language is Spanish.

with highest weights. However, those words in the Gold-Standard happen to be among the most probable translations of the GIZA++ dictionary, and when the combination between both values is performed, the correct translations ("ensayo", "experimento" and "prueba"), appear in the final output of the system. Hence, the solution given for this particular case is successful, although it represents an example of the case in which most of the useful information comes from the translation probabilities given by GIZA++.

In the second example, we can also observe a case in which a successful solution is achieved. In this case, the CO-Graph system alone is able to propose most of the correct translations for the target word in the test sentence. Words "presión" and "carga", present in the Gold-Standard, obtain the highest weights from the CO-Graph system alone. The translation with highest probability according to GIZA++ ("cepa") is not among those contained in the Gold-Standard. This word has a small weight in the CO-Graph system ranking, and hence when the combination of weights is performed to obtain the final output, this word does not get a high value in the final ranking. Accordingly, a possible decrease of the performance, that could be caused by given too much importance to that word, is avoided thanks to the weights given by the CO-Graph system alone. Apart from this, the effect of the translation probabilities can be seen for the word "tensión", which obtains a slightly smaller CO-Graph weight, and thanks to the translation probability given by GIZA++ is proposed to be the second most probable translation in the final solution.

4.7.3 Parameter Adjustment

When the graph is constructed, a threshold for the p-value p has to be set, in order to indicate the highest value of p for which the number of co-occurrences of two words is considered to be statistically significant and therefore a link is created between them. As this threshold decreases, the graph becomes more restrictive, and hence the number of edges also decreases. We have used the trial dataset provided in the SemEval 2010 competition for analysing the influence of the threshold in an exhaustive way. Then, based on those results we will select a specific threshold for performing the experiments in all the languages and all the tested algorithms, on the 2010 and 2013 test datasets.

Figure 4.10 shows the evolution of the behaviour of the system as we decrease the threshold for building the graph, from $p = 10^{-5}$ to $p = 10^{-17}$.

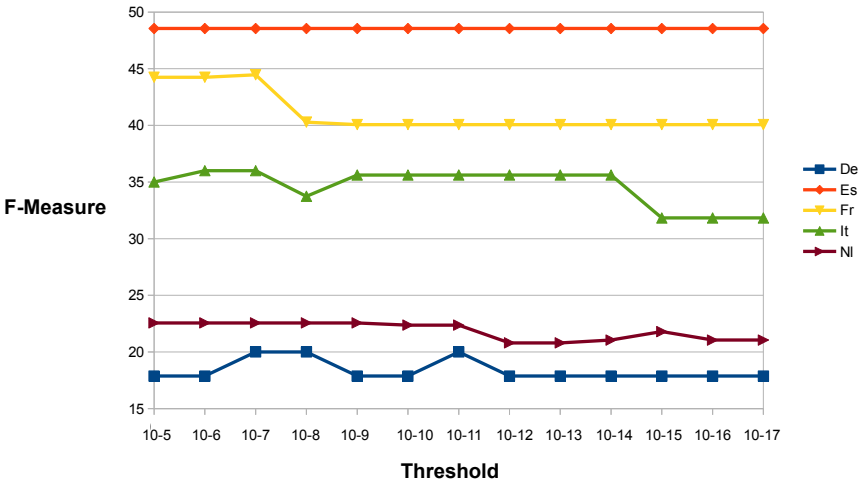


Figure 4.10: Evolution of the F-Measure achieved in a single experiment, as the threshold decreases (graph becomes more restrictive). Results for the PageRank algorithm in the "Complete graph approach" using the trial dataset of the 2010 SemEval competition.

As we can observe in the figure, for all the languages best results are achieved with values of the threshold between $p = 10^{-5}$ and $p = 10^{-11}$, while smaller thresholds lead to a decrease of the performance of the algorithm. Within this smaller range of thresholds, the F-Measure values are quite similar. This can be observed in most of the experiments conducted in this work, hence selecting any value of the threshold in that range would provide similar results. In particular, we have selected a threshold value of $p = 10^{-6}$ for all the experiments in this section, given any language or algorithm. This way, we want to test the robustness of our system under the same conditions that the systems participating in the SemEval competitions.

4.7.4 Baseline: Most Frequent Sense

As we stated above, GIZA++ not only generates a dictionary from a language to another in an automatic way, but it also provides the probability of each translation of each word. Hence, the use of GIZA++ for creating the bilingual dictionaries allows us to generate a baseline to which compare our system: the most probable translations (Most Frequent Sense or MFS). As we are considering two different evaluation schemes, we will provide the most probable translation for the "Best Evaluation" scheme, and the five most probable translations for the "Out-Of-Five Evaluation" scheme. Table 4.3 illustrates the results offered by a MFS approach applied to the 2010 and 2013 SemEval test datasets. Both evaluation schemes ("Best" and "Out-Of-Five") are shown in the table.

		De	Es	Fr	It	Nl
SemEval 2010	**Best Evaluation**	12.07	16.11	19.63	15.83	13.89
	OOF Evaluation	25.29	44.02	44.58	40.55	37.11
SemEval 2013	**Best Evaluation**	15.41	19.81	23.97	19.95	18.94
	OOF Evaluation	32.89	49.75	50.97	49.71	43.35

Table 4.3: MFS Baseline. Results (F-Measure in %) obtained by a MFS approach for the Best and Out-Of-Five (OOF) evaluation schemes in the SemEval 2010 and SemEval 2013 competitions.

As expected, the results for the "Best" evaluation scheme are always lower than those for the more relaxed "Out-Of-Five" scheme. In general, Spanish, French and Italian obtain better results than Dutch and German, and this trend will be repeated along all the experiments performed with our system, and also along the results offered by different systems participating in the competitions. The fact that Dutch and German are languages for which disambiguation is a more difficult task, can be caused by the higher number of translations, in general, that they present for each target word, as we observed in Table 4.2. It is important to state that this baseline based on most frequent senses is usually hard to overcome by systems participating in CLWSD tasks. Actually, the baselines proposed by the organizers of the 2010 and 2013 competitions, which were also based on a MFS approach, were not surpassed by many of the participating systems, even supervised ones. Moreover, when the baselines were outperformed, the differences were small. This fact will be shown in following sections containing results from the best systems participating in the competitions.

4.7.5 Word-Based Graphs versus Complete Graphs

In this section we present the results obtained by our system, using the approaches described in Section 4.5.2, this is, Word-Based Graphs and Complete Graphs, and we compare both

techniques. Table 4.4 shows the results for the Word-Based Graph and Complete Graph approaches in the SemEval 2010 test dataset, and Table 4.5 shows the results obtained by both approaches in the SemEval 2013 test dataset. All the algorithms described in Section 4.6 are tested and compared. Results shown correspond to the F-Measure value.

Word-Based graph approach: SemEval 2010 Competition						
		De	**Es**	**Fr**	**It**	**Nl**
Best	**Walktrap**	**12.33**	18.89	**20.96**	15.95	**13.88**
	Chinese Whispers	11.97	18.90	19.89	**16.57**	13.56
	Dijkstra	12.29	18.05	20.95	16.17	13.86
	Basic PageRank	10.71	18.99	20.49	15.72	12.36
	PageRank with Priors	9.44	**19.47**	19.90	16.51	11.92
OOF	**Walktrap**	**25.75**	**47.04**	**46.85**	**41.76**	36.45
	Chinese Whispers	25.51	46.58	46.15	41.71	36.46
	Dijkstra	25.53	46.92	46.13	41.66	**36.50**
	Basic PageRank	25.29	45.20	45.09	41.73	35.56
	PageRank with Priors	24.71	45.34	45.40	41.10	36.39

Complete graph approach: SemEval 2010 Competition						
		De	**Es**	**Fr**	**It**	**Nl**
Best	**Walktrap**	**13.30**	**19.07**	20.46	15.38	12.53
	Chinese Whispers	13.23	18.98	20.19	15.60	13.13
	Dijkstra	13.23	18.98	**20.89**	16.25	**14.39**
	Basic PageRank	10.67	19.01	20.69	14.76	12.49
	PageRank with Priors	11.42	19.06	20.11	**16.34**	12.53
OOF	**Walktrap**	27.52	47.09	**47.53**	**41.60**	35.87
	Chinese Whispers	26.89	46.51	46.69	41.34	35.65
	Dijkstra	26.91	**47.32**	46.50	**41.60**	36.07
	Basic PageRank	**27.54**	46.43	45.40	40.08	36.23
	PageRank with Priors	26.98	46.67	46.49	40.93	**36.92**

Table 4.4: Results (F-Measure in %) obtained by the Word-Based Graph and Complete Graph approaches of the CO-Graph system for the Best and Out-Of-Five (OOF) evaluation schemes in the SemEval 2010 competition. Results for the five different disambiguation algorithms are presented, for the five languages involved in the task. Bold highlights the algorithm that reaches the best results for each approach, language and evaluation scheme.

Regarding the different disambiguation algorithms, as we can observe in the tables, in general the techniques based on the Walktrap algorithm and on Dijkstra's algorithm show the best results, both for the "Best" and the "Out-Of-Five" evaluation schemes. There are some cases in which the other community-based algorithm, Chinese Whispers, slightly outperforms those results. However, in those cases, the improvement with respect to the second best result is really small. Also, the PageRank with Priors algorithm, and the Basic PageRank algorithm obtain the best result for some of the experiments. In these cases, the improvement

Word-Based graph approach: SemEval 2013 Competition						
		De	**Es**	**Fr**	**It**	**Nl**
Best	**Walktrap**	16.17	22.71	**25.09**	20.88	16.91
	Chinese Whispers	15.63	22.71	24.53	21.20	17.13
	Dijkstra	**16.24**	22.21	24.97	20.98	**17.60**
	Basic PageRank	13.57	23.27	24.00	20.27	14.88
	PageRank with Priors	11.35	**23.75**	22.59	**21.22**	15.56
OOF	**Walktrap**	**32.53**	**52.89**	51.54	51.53	42.55
	Chinese Whispers	32.39	52.22	**52.00**	**51.77**	**42.74**
	Dijkstra	31.92	52.62	51.92	51.04	42.67
	Basic PageRank	31.18	50.03	49.67	50.45	41.93
	PageRank with Priors	31.17	49.37	49.41	49.40	41.15

Complete graph approach: SemEval 2013 Competition						
		De	**Es**	**Fr**	**It**	**Nl**
Best	**Walktrap**	17.81	22.96	25.06	20.26	15.88
	Chinese Whispers	**17.91**	**23.33**	25.03	20.46	16.84
	Dijkstra	17.75	22.89	**25.45**	**21.02**	**17.96**
	Basic PageRank	12.66	22.40	24.47	19.66	14.56
	PageRank with Priors	13.54	22.53	22.75	20.05	15.54
OOF	**Walktrap**	35.44	**52.80**	**52.26**	**51.77**	42.67
	Chinese Whispers	34.75	52.07	51.75	51.63	42.60
	Dijkstra	34.84	52.56	52.04	51.57	43.42
	Basic PageRank	**35.53**	50.16	49.94	47.67	43.36
	PageRank with Priors	34.28	49.95	50.77	49.24	**43.49**

Table 4.5: Results (F-Measure in %) obtained by the Word-Based Graph and Complete Graph approaches of the CO-Graph system for the Best and Out-Of-Five (OOF) evaluation schemes in the SemEval 2013 competition. Results for the five different disambiguation algorithms are presented, for the five languages involved in the task. Bold highlights the algorithm that reaches the best results for each approach, language and evaluation scheme.

is also not significant. Hence, we can assume that algorithms based on distances and paths between translations of words from the context and potential translations of the target word (community-based and Dijkstra) perform better than those based only on the outdegree of the nodes and random jumps (PageRank). The community-based algorithms take into account the weights of the links in the co-occurrence graph for extracting the communities, while Dijkstra's algorithm considers them (their inverse) for obtaining the shortest path between two nodes. Therefore, an accurate weighting of the connections of the nodes in the graph is important for obtaining good results.

The comparison between the Word-Based Graph and the Complete Graph approaches shows

a general trend in which results obtained by the Complete Graph approach are better than those provided by the Word-Based Graph approach, although most of the improvements are small. This can be due to the greater completeness of the general graph used for disambiguation, that takes into account not only information about the target word, but about the structure of the corpus in a more generalistic way. This improvement is particularly clear when it comes to German, Spanish and French. Accordingly, and considering that a unique graph for disambiguating all the target words is better in terms of efficiency, we will select the Complete Graph approach and the technique based on Dijkstra's Algorithm for the rest of the evaluation, and for the comparison with other state-of-the-art systems.

Once that we have selected a particular threshold value and a particular technique for building the graphs and performing the disambiguation, we want to compare its results with those obtained by the same technique (Dijkstra's algorithm with complete graphs and a threshold value of 10^{-6}), but without the combination with the translation probabilities given by GIZA++. We examine the performance of the CO-Graph system alone, and determine if the combination proposed in Section 4.6.4 improves the original results. Table 4.6 shows this comparison.

SemEval 2010 Competition						
		De	**Es**	**Fr**	**It**	**Nl**
Best	**CO-Graph**	5.56	8.57	7.22	6.15	6.53
	CO-Graph + GIZA++	**13.23**	**18.98**	**20.89**	**16.25**	**14.39**
OOF	**CO-Graph**	26.00	41.25	38.66	33.77	29.57
	CO-Graph + GIZA++	**26.91**	**47.32**	**46.50**	**41.60**	**36.07**

SemEval 2013 Competition						
		De	**Es**	**Fr**	**It**	**Nl**
Best	**CO-Graph**	7.43	7.91	6.55	7.87	7.17
	CO-Graph + GIZA++	**17.75**	**22.89**	**25.45**	**21.02**	**17.96**
OOF	**CO-Graph**	33.08	46.10	38.90	39.59	33.35
	CO-Graph + GIZA++	**34.84**	**52.56**	**52.04**	**51.57**	**43.42**

Table 4.6: Comparison of results (F-Measure in %) for the selected configuration of the system (threshold value $p = 10^{-6}$, Dijkstra's algorithm, Complete Graph approach), between the CO-Graph system alone and the CO-Graph system combined with the prior translation probabilities given by GIZA++, for the 2010 and 2013 test datasets, in both the "Best" and "Out-Of-Five" evaluation schemes. Bold highlights the technique that reaches the best results.

The table clearly shows that the original CO-Graph system is substantially improved by combining its output with the translation probabilities given by GIZA++. Specifically, results for the "Best" evaluation scheme present the highest improvements. This suggests the importance of the backoff provided by the most frequent sense of a given target word for selecting only the most suitable translations. In the "Out-Of-Five" scheme, important improvements are also achieved in most of the languages. This confirms our hypothesis

about the benefits derived from the use of the prior translation probabilities given by the bilingual dictionaries, suggested in Section 4.6.4. Nevertheless, the results obtained by CO-Graph alone are still competitive, specifically in some cases of the "Out-Of-Five" evaluation. The system that combines the CO-Graph weights and the translation probabilities will be used along the rest of the chapter for performing comparisons with other state-of-the-art systems.

4.7.6 Comparative

In this section, Table 4.7 shows the comparative between our system and the best unsupervised systems participating in the SemEval 2010 and 2013 competitions, described in Chapter 2. The MFS baseline proposed in Section 4.7.4 is also included in the table for comparison. Also, the multilingual system described in (Navigli and Ponzetto, 2012) is compared separately, since it did not participate in the competitions, and only proposes results for the "Best" evaluation scheme of the SemEval 2010 test dataset.

			De	Es	Fr	It	Nl
SemEval 2010	**Best**	**CO-Graph**	13.23	18.98	20.89	16.25	**14.39**
		Best System	13.71	19.68	21.84	15.47	10.63
		Multilingual	**18.26**	**23.65**	**24.61**	**19.05**	N/A
		MFS	12.06	16.11	19.63	15.83	13.89
	OOF	**CO-Graph**	26.91	**47.32**	46.50	**41.60**	**36.07**
		Best System	**33.01**	35.65	**49.20**	40.52	21.37
		Multilingual	N/A	N/A	N/A	N/A	N/A
		MFS	25.29	44.02	44.58	40.55	37.11(*)
SemEval 2013	**Best**	**CO-Graph**	**17.75**	22.89	**25.45**	21.02	**17.96**
		Best System	8.13	**32.16**	24.56	**21.20**	9.89
		Multilingual	N/A	N/A	N/A	N/A	N/A
		MFS	15.41	19.81	23.97	19.95	18.94(*)
	OOF	**CO-Graph**	**34.84**	**52.56**	**52.04**	**51.57**	**43.42**
		Best System	23.71	49.01	45.37	40.25	27.11
		Multilingual	N/A	N/A	N/A	N/A	N/A
		MFS	32.89	49.75	50.97	49.71	43.35

Table 4.7: Comparative of the results (F-Measure in %) obtained by the CO-Graph system, the unsupervised systems obtaining the best results, and the MFS approach, for the SemEval 2010 and SemEval 2013 competitions, in the five proposed languages. Bold highlights the best system (CO-Graph, best unsupervised or multilingual approach), and asterisk (*) indicates the cases in which the CO-Graph system has not been able to overcome the MFS approach.

The table shows that the CO-Graph system overcomes the best unsupervised systems that participated in the SemEval competitions in most of the cases (13 out of 20). The multilingual

system, although it did not participate in the competitions, outperforms our system and the best unsupervised systems in all the cases for which it proposes results ("Best" scheme for German, Spanish, French and Italian). However, although unsupervised, this system uses external resources such as BabelNet (Navigli and Ponzetto, 2010) and WordNet (Fellbaum, 1998) for performing the disambiguation, while our system only makes use of a parallel corpus for extracting the knowledge. In general, our system performs better for the "Out-Of-Five" evaluation scheme, compared to the "Best" evaluation, and also for the 2013 test dataset, compared to the 2010 dataset. In the 2013 edition all the unsupervised systems are surpassed in the "Out-Of-Five" scheme, and only one of them, called NRC-SMT (Carpuat, 2013), clearly overcomes our system in the "Best" scheme, for the Spanish language. This system is focused on the Spanish language only, so its coverage of the problem is lower than ours. In the Italian language, a system, called LIMSI (Apidianaki, 2013) obtains slightly better results than ours, but the difference is so small (an improvement of 0.18%), that it can be considered as a tie. The system that overcomes ours in the "Out-Of-Five" scheme of the 2010 dataset (German and French) is called T3-COLEUR (Guo and Diab, 2010). Although it is an unsupervised system, it takes additional information from some external resources, since WordNet synsets are used to augment the translation correspondences and yield more translation variability.

Our system is able to overcome the MFS approach in most of the cases. Only in two of the 20 cases the MFS approach presents better results, but the differences are really small (around 1%).

4.8 Conclusions

In this chapter we have presented an approach to perform Cross-Lingual Word Sense Disambiguation (CLWSD), based on the construction of the co-occurrence graph, which contains the knowledge of a corpus in a structured way. We have shown the validity of the unsupervised graph-based technique, which uses the whole document as a coherent piece of information, while other works consider windows of a specific size for building the context and calculating the co-occurrences. The results obtained support our hypothesis about the effectiveness of the approach. Results also show that algorithms based on the weights of the links in the co-occurrence graph (community-based techniques and Dijkstra's algorithm) performs slightly better on average that the other proposed techniques. The community-based techniques use the weights of the links in the co-occurrence graph for extracting the communities, and the technique based on Dijkstra's algorithm takes those same weights into account for finding the shortest paths between nodes. Results indicate that the weights assigned to the links in the co-occurrence graph construction process are useful for determining the influence of the surrounding words of the target word. However, the

differences between those algorithms and the algorithms based on the relative importance of each node in the graph (PageRank, in its different variants), are not very large.

We have shown the relevance of considering the probabilities of occurrence of the different translations taken into account for a target word, through the combination of weights proposed by the system and the values of these translation probabilities. The conclusion that can be drawn from this fact is that information regarding the possible translations of the target words (which can be obtained in an automatic, unsupervised way), is a key knowledge for systems performing CLWSD, and hence a bilingual dictionary offering this kind of information is very important for building competitive systems.

Results also show that our system is among the best unsupervised systems that participated in SemEval competitions, overcoming all of those systems for some languages, and obtaining the second best results for some others. In general, we can observe that results of the "Out-Of-Five" evaluation scheme are better than those achieved using the "Best" evaluation scheme. In fact, our system overcomes all the participating unsupervised systems in the SemEval 2013 competition, for the "Out-Of-Five" evaluation scheme. The difference of performance for the two evaluation schemes may be due to the fact that the "Best" evaluation requires more pronounced differences among the weights. This could be achieved by increasing the influence of the context.

It is also important to indicate that those unsupervised systems that overcome our results make use of some additional external resources, or are focused only on one language, as we stated above, while our system only use the proposed corpus for extracting knowledge, and provides good results for all the languages proposed in the tasks. Hence, we can conclude that the performance of the system is better than that presented by other unsupervised systems using the same resources. Our system also presents a remarkable robustness and coverage of the problem.

Regarding the considered languages, the best results are obtained for Spanish, French and Italian, while the lowest ones correspond to German, which has been proved to be a more difficult task to solve in both competitions for all the participating systems. The combination of the output of the system and the most probable translations provided by the dictionary overcomes the results obtained by the baseline (MFS approach) in most of the cases. As it can be seen in the results of the 2010 and 2013 competitions (even for supervised systems), the fact of outperforming the MFS approach is not a small achievement, as this kind of baselines are usually hard to overcome in CLWSD tasks. In general, we can point out the robustness achieved by our system, with respect to parameters (such as the p-value) and languages.

5

DICTIONARIES FOR CROSS-LINGUAL WORD SENSE DISAMBIGUATION

What's another word for Thesaurus?

Steven Wright

Contents

5.1 Introduction

Many issues arise along the disambiguation process, the choice of an adequate bilingual dictionary being one of the most important for ensuring the good performance of a system. Chapter 4 introduced Cross-Lingual Word Sense Disambiguation as the task of automatically determining the contextually appropriate translation for a given word, from a source language to a target one. As we briefly mentioned in Section 4.1, in this kind of tasks the disambiguation process usually relies on the existence of a bilingual dictionary which contains the set of potential translations of any ambiguous word, which eventually represents the ambiguity to be solved. Hence, the bilingual dictionary that provides translations, both for words surrounding the target word (context) and for the target word itself, is a key part of the disambiguation process. This dictionary offers the potential translations of the target word, and any system which performs the disambiguation has to choose, among the translations, those which are considered most suitable for the particular sentence. This selection is then matched against an expected output or Gold-Standard to determine a score for that specific test instance.

In this chapter, we intend to compare the use of bilingual dictionaries of different nature: manually created by experts, semi-automatic, i.e. extracted with automatic tool but with human supervision or intervention, collaboratively edited by different authors, and statistical dictionaries. As it has been shown before, this last type of dictionaries, automatically created without human supervision, provide a much larger number of translations, at the price of introducing noise. However, apart from their size and the coverage they can present (denoted by the number of different translations for each word), statistical dictionaries provide information about the translation probabilities, since their construction is based on statistical characteristics. The other dictionaries studied in this chapter do not usually present this kind of information. Considering that CLWSD tasks are based on translations of words used in general sentences, we can expect information about the most frequent translations to be useful for our purposes.

We will analyse different dictionaries that provide the candidate translations, and compare the results obtained using them, both in ideal conditions, and inside the particular unsupervised CLWSD system (CO-Graph) presented in the previous chapter. These results will give us important insights about the potential variations of the effectiveness of the CLWSD system according to the choice of the bilingual dictionary.

We will focus on the English-Spanish cross-lingual disambiguation, and on the out-of-five evaluation proposed in both SemEval tasks previously used: task 3 from SemEval 2010 and task 10 from SemEval 2013. As we already mentioned, this evaluation scheme requires the systems to provide up to five guesses for each target word in each context, without penalising them due to the number of guesses.

The first objective of this chapter is the design of some experiments to compare different dictionaries in a general framework of a disambiguation task. For this purpose, we have developed a frame in which theoretical limits can be found for the performance of each of the analysed dictionaries for well defined CLWSD tasks. Once that we find these limits (upper bounds), we intend to explore the actual performance of a particular CLWSD system in the task, and analyse its results depending on the dictionary.

The rest of the chapter is organized as follows: Section 5.2 describes the configuration of the CO-Graph system, which is used through the rest of the chapter to compare the different dictionaries. Section 5.3 explains in detail the different considered dictionaries. The characteristics of the evaluation framework used for testing the dictionaries are shown in Section 5.4. Section 5.5 analyses the results that could be achieved by an ideal system, depending on the bilingual dictionary used. An error analysis concerning those results is conducted in Section 5.6. In Section 5.7, the dictionaries are tested within the CO-Graph system previously introduced, and the obtained results are analysed. Section 5.8 offers a comparative between these CO-Graph results and other systems participating in the SemEval competitions. Finally, conclusions are detailed in Section 5.9.

5.2 Configuration of the CO-Graph System

A system will be used in this chapter for testing the performance of the proposed bilingual dictionaries in real conditions, that is, we will embed the different dictionaries within a particular system performing Cross-Lingual WSD. This system is a particular case of the CO-Graph system presented in Chapter 4. The base of knowledge for the disambiguation system is the Europarl parallel corpus (Koehn, 2005). The co-occurrence graph will be created using the same procedure described in Chapter 3.

In the previous chapter we have studied different disambiguation algorithms which performed the last step of our system were analysed. However, in this case we want to study the impact of using different dictionaries, hence we will select a disambiguation algorithm which will be used throughout the experiments in this chapter. In particular, we will apply the technique based on community detection. Communities will be located inside the graph through the use of the Walktrap algorithm (Pons and Latapy, 2005), and a community graph will be created from this original graph. This community graph links clouds of words, each one of them containing related words, in terms of co-occurrence. Finally, the translations of words (in this case, nouns) of the context and the potential translations of the target word will be found inside the community graph, and the distances between communities containing translations of the target word and communities containing translations of words of the context will be calculated.

Figure 5.1 illustrates the complete CO-Graph system, with all its phases: the extraction of words from the test instance, the translation of those words, the construction of the co-occurrence graph and the community graph, and finally the disambiguation step, involving the community graph and the translated words.

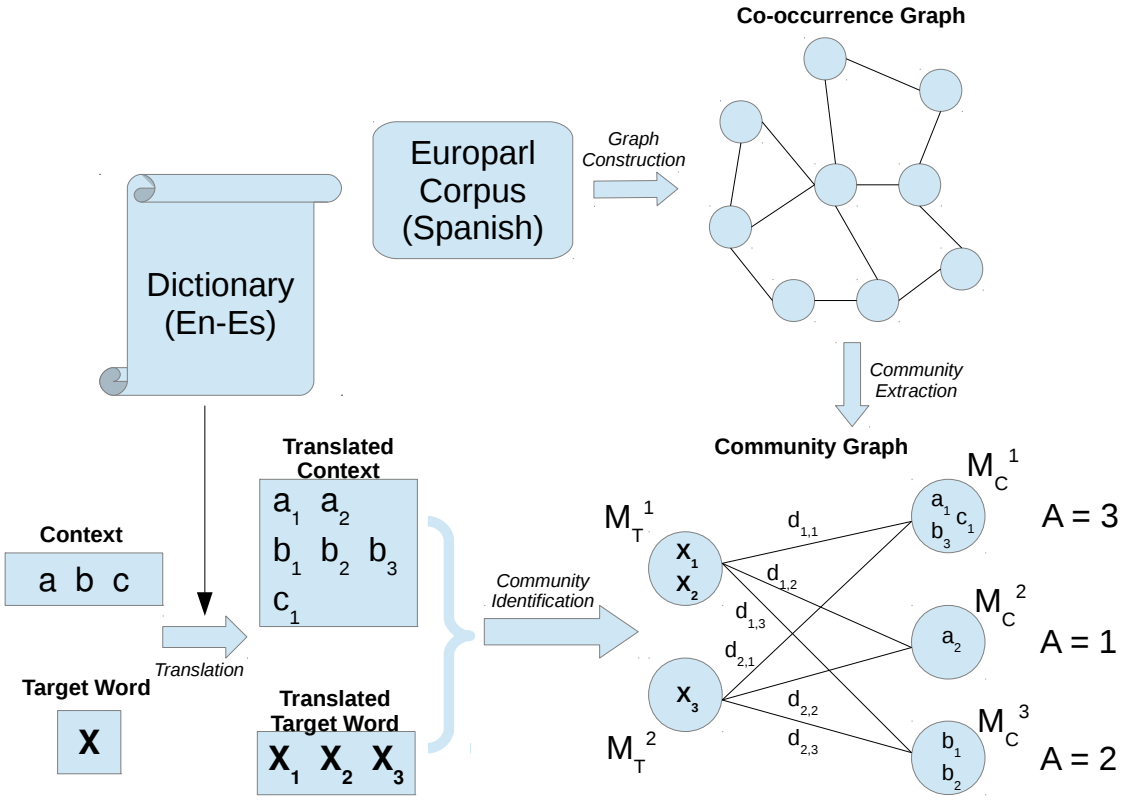


Figure 5.1: Diagram and example of the CLWSD system. The community graph is extracted from the co-occurrence graph, and used to compute the distances between words from the context and the target word. Communities named with "M_T" contain translations of the target word, and communities named with "M_C" contain translations of the words of the context. The letter "A" represents the number of translations from words of the context that can be found in each of the "M_C" communities.

5.3 Bilingual Dictionaries

In this section, we present the main characteristics of the different bilingual dictionaries considered for our purposes. They are four English-Spanish dictionaries: a manually created dictionary, built by experts, which will be denoted as "external dictionary" along the rest of the chapter, a collaboratively edited dictionary, a semi-automatic dictionary and a statistical, automatically created parallel corpus-based dictionary. All of them are described below.

- **External dictionary**: This dictionary (López-Ostenero, 2002) is completely external to the main task. It is a generic bilingual dictionary, which has no relation to the source

of knowledge in the task (the Europarl corpus). The results offered by this dictionary, both for the ideal system and for the CO-Graph system, are considered to be a baseline, and hence a goal of the other dictionaries is to overcome those results.

- **MCR dictionary**: The Multilingual Central Repository (Atserias et al., 2004) is a lexical knowledge base (LKB) that constitutes a multilingual large scale linguistic resource for many semantic processes, due to the amount of multilingual knowledge that it contains (Agirre and Soroa, 2008). This LKB contains lexical information about five different languages: English, Spanish, Basque, Catalan and Galician, and is based on the WordNet and EuroWordNet projects. Synsets from different languages are linked through the Inter-Lingual Indices (ILIs). From the ILIs present in MCR 3.0 (Gonzalez-Agirre et al., 2012), we have extracted direct translations from English to Spanish to create our bilingual dictionary.

- **BabelNet dictionary**: BabelNet (Navigli and Ponzetto, 2010) is a very large semantic multilingual network that links Wikipedia information to WordNet synsets in an automatic way. The whole resource could be considered as a semi-automatic dictionary, since multilingual information comprises both manual translations from Wikipedia, although automatically mapped to WordNet information, and translations obtained by applying machine translation to the SemCor corpus (Miller et al., 1993). For any word in the English language, we can obtain all the possible senses of the word, and their corresponding translations in the final language (in our case, Spanish).

- **GIZA++ dictionary**: The statistical aligner GIZA++ is able to extract one-to-many translations from a target word and their corresponding probabilities of occurrence. For this aim, it uses a parallel corpus as knowledge base, in our case the Europarl corpus. In the first step, the GIZA++ tool performs a word alignment over the initial corpus, without any preprocessing. Once that the alignment is done, we obtain a probability table. This table links every word in the original language (in this case, English) to each of its possible translations in the final language (in this case, Spanish), and assigns a probability of occurrence. Due to the automatic and statistical nature of the algorithm implemented by GIZA++, the number of translations that it proposes for each English word is very high. This fact may introduce noise in the translation process so a technique to reduce this inducted noise and thereby improve the accuracy is needed. For this purpose, we performed the alignment in the other direction, i.e., obtaining a one-to-many word alignment from Spanish to English, and then calculated the intersection of both probability tables. In this way, we obtain an English-Spanish dictionary, ensuring that every English-Spanish translation has an equivalent Spanish-English translation. We have excluded stop words for building the dictionary.

Table 5.1 shows some statistics about the dictionaries used in this chapter. Specifically, we can observe the number of entries, maximum number of translations presented by a word, and the average number of translations for all the words in the dictionary.

	Entries	Max # translations	Average # translations
External	50,911	87	2.32
MCR	35,440	56	2.09
BabelNet	384,832	89	2.62
GIZA++	34,815	1,344	7.51

Table 5.1: Statistics from the bilingual dictionaries. Column "Entries" represents the total number of entries of the dictionary. Column "Max # translations" shows the maximum number of translations for a word. Column "Average # translations" shows the average number of translations in the complete dictionary.

Regarding the number of entries in the dictionary, we can observe that the BabelNet dictionary presents many more words than any other dictionary. This can be due to the completeness of the dictionary, which can be considered more as an encyclopaedic dictionary, since not only synsets from WordNet, but also entities from Wikipedia, are collected to build it. However, the total number of entries is not important in this case, given that all the words in the test sentences are covered by all the dictionaries. The average number of translations is a more important fact when we want to analyse the impact of each dictionary. In this case, we can observe that most of the dictionaries offer an average number of translations between 2 and 3. Nevertheless, the GIZA++ dictionary offers many more translations per word than the other dictionaries. This can lead to a wider coverage of the problem. On the other hand, and regarding a real system, this fact may imply a drawback, considering that a high number of possible translations for a target word could prevent the system from finding the most suitable ones. That is, the coverage would be high, but the precision may decrease.

5.4 Datasets and Evaluation

In this chapter, we will consider the same evaluation framework used in Chapter 4, based on the one proposed in task 3 of SemEval 2010 and task 10 of SemEval 2013 competitions. As stated before, the evaluation scheme will be "out-of-five", in which the system has to select five of the potential translations for each test instance. We use the F-Measure value for illustrating the results achieved.

Regarding the datasets, Table 5.2 offers more information about the statistics of the dictionaries, focused on the 20 words composing the datasets. More specifically, it shows the number of translations offered by each dictionary for each possible target word in the test datasets.

The table clearly shows the differences in number of translations for each target word depending on the bilingual dictionary. We can observe that the external dictionary, the

Word	External	MCR	BabelNet	GIZA++
coach	15	13	27	**8**
education	6	4	10	52
execution	4	6	14	30
figure	29	25	25	146
job	17	14	28	133
letter	**3**	4	6	46
match	15	26	18	101
mission	6	7	8	35
mood	4	**3**	**4**	32
paper	10	8	12	64
post	30	21	11	72
pot	**43**	**41**	**80**	21
range	25	17	30	100
rest	22	11	13	87
ring	31	13	21	34
scene	15	9	19	46
side	19	15	26	**191**
soil	10	5	10	10
strain	31	13	32	48
test	15	7	7	89
Mean	**17.50**	**13.10**	**20.05**	**67.25**

Table 5.2: Number of translations of the words in the datasets, for each dictionary: External (second column), MCR (third column), BabelNet (fourth column) and GIZA++ (fifth column). Bold represents maximum and minimum values for each dictionary.

dictionary based on MCR and the dictionary based on BabelNet present similar behaviour. In the three cases, the word which presents the highest number of translations is "pot", while the word "mood" presents the lowest number of translations for the MCR and BabelNet dictionaries, and the second lowest for the external dictionary. On the other hand, the behaviour of the GIZA++ dictionary is completely different, as the word presenting the highest number of translations is "side" and the word presenting the lowest number is "coach". These differences can be due to the automatic nature of the dictionary generated with GIZA++. The other dictionaries present human intervention in their construction, which can lead to a different number of translations. Apart from this fact, it is important to notice the high number of possible translations produced in the GIZA++ dictionary, which may lead to decrease the performance. To avoid this decrease, we also considered a restricted GIZA-based dictionary, with a maximum of ten possible translations per word. These ten translations are those that present the highest probabilities of occurrence. Some experiments regarding the value of maximum translations per word have been done, showing that a pruning value of ten translations per word provides the best results. This dictionary will be denoted as "GIZA10" along the rest of the chapter.

5.5 Influence of the Dictionaries on an Ideal System

A good indicator for understanding how the dictionary can modify the performance of a system in a CLWSD task is the highest score that could be achieved by a perfect system for a given dictionary. In this particular case, we define the upper bound for a given dictionary as the best result that a system that uses this dictionary can achieve, according to the Gold-Standard. Since we are working with datasets from two past competitions, we have access to the Gold-Standards used for the evaluation. Then, for building the best guessing that a system could give, we take for every context of every target word those translations from the dictionary that are also in the solution provided by the Gold-Standard. If there are words in the Gold-Standard for this context that are not present in the dictionary, random words are selected to complete the requested five word guessing. In the proposed dictionaries we do not take into account those translations that contain more than one word.

Tables 5.3 and 5.4 show the highest F-Measure, for each word in average, that can be achieved by any system using the five considered dictionaries. The last column represents an upper bound obtained by applying the same process to the Gold-Standard itself, but excluding from the proposed solution those translations containing more than one word, since the co-occurrence graph used in CO-Graph only considers one-word translations (nodes of the graph represent one single word). More particularly, Table 5.3 shows the results for the 2010 test set, and Table 5.4 the results for the 2013 test set.

The dictionary obtained with GIZA++ and without restrictions (Column **GIZA**) is the resource that would allow an ideal system to obtain the best results. However, due to the noise that the high number of translations of the dictionary induces, in the rest of the chapter we will use GIZA10. In the tables we can also observe that the dataset for 2013 ideally allows the systems to achieve better results, as the upper bounds are higher in all cases. The last column, representing the modified Gold-Standard (without translations containing more than one word), gets close to a perfect performance. However, its accuracy is not 100% due to the mentioned exclusion of multi-word translations. Hence, it provides some clues about the reduction of accuracy due to this exclusion. There are some words for which the external dictionary obtains a higher upper bound than the GIZA++ dictionary ("*post*" and "*pot*"). This may be due to the specific characteristics of those words (number of translations, differences between translations, ...). Overall, most of the words present significant potential improvements in their upper bounds when we use the GIZA++ dictionary. A deeper analysis regarding the words which present better performance with the other dictionaries is done in Section 5.6. Comparing Tables 5.3 and 5.4 with Tables 5.1 and 5.2 we observe a direct correlation between the translation average in a dictionary and the performance (average F-Measure) of an ideal system using that dictionary. As the upper bounds are basically representing the coverage of each dictionary (the maximum performance that could be achieved), this correlation is expected: as the number of possible translations increases, the

Upper Bounds 2010						
Word	**ExtDic**	**MCR**	**BabelNet**	**GIZA10**	**GIZA**	**Gold**
coach	63.17	58.31	**76.89**	**76.89**	**76.89**	96.60
education	77.82	77.82	80.88	84.13	**94.00**	98.19
execution	53.26	53.26	62.94	67.77	**80.00**	89.35
figure	46.97	44.27	49.02	62.63	**84.90**	95.03
job	54.38	31.55	53.01	61.58	**74.10**	83.02
letter	37.51	37.51	40.94	42.68	**57.66**	93.19
match	46.74	55.79	55.79	26.41	**71.80**	99.71
mission	55.06	55.06	55.06	56.19	**76.12**	99.18
mood	14.20	23.27	26.42	62.32	**68.97**	77.64
paper	39.45	25.41	28.08	43.33	**64.92**	97.69
post	47.27	37.28	**49.46**	16.94	39.30	83.57
pot	**55.15**	32.37	45.57	38.60	48.71	89.70
range	17.66	15.15	21.29	17.96	**45.44**	84.77
rest	30.90	33.27	34.85	26.08	**36.48**	89.73
ring	42.04	29.00	30.49	50.65	**66.86**	98.83
scene	42.46	42.46	46.88	61.44	**80.86**	90.08
side	40.55	33.26	36.30	43.28	**70.43**	84.98
soil	63.06	63.06	73.69	**98.07**	**98.07**	99.27
strain	26.55	26.55	39.02	67.07	**83.17**	93.41
test	68.92	59.11	66.38	80.20	**87.00**	95.22
Mean	**46.16**	**41.69**	**48.65**	**54.22**	**70.28**	**91.97**

Table 5.3: Upper bounds (F-Measure in %) for SemEval 2010 test dataset, obtained with different translation dictionaries: external dictionary (column **ExtDic**), dictionary based on the Multilingual Central Repository (column **MCR**), BabelNet-based dictionary (column **BabelNet**), complete GIZA++ dictionary (column **GIZA**) and pruned GIZA++ dictionary (column **GIZA10**). Last column represents results obtained by the Gold-Standard without considering multi-word translations. Bold represents best results for each word without taking the Gold-Standard into account.

probability of covering more words from the Gold-Standard is higher, and hence the ideal performance of the system also increases.

Figure 5.2 shows an example of the process of construction of the upper bounds for any dictionary. Given a sentence and its Gold-Standard, we extract from the dictionary those words (highlighted in bold letters in the example) that appear in the Gold-Standard. The rest of the words, up to five, are randomly selected from those proposed by the dictionary. In the example, the external, MCR and BabelNet dictionaries contain two words appearing in the Gold-Standard ("escena" and "panorama"). On the other hand, the GIZA10 dictionary contains three coincident words ("ámbito", "escena" and "panorama"). Hence, an ideal system based on GIZA would obtain a better result for this particular instance.

Upper Bounds 2013						
Word	**ExtDic**	**MCR**	**BabelNet**	**GIZA10**	**GIZA**	**Gold**
coach	76.50	73.53	**83.83**	**83.83**	**83.83**	100.00
education	77.17	76.83	75.34	83.83	**88.98**	92.67
execution	50.29	50.29	65.53	61.48	**75.81**	86.68
figure	57.49	52.77	56.00	69.55	**88.83**	99.53
job	66.93	40.54	56.99	63.51	**76.54**	84.34
letter	59.06	59.06	60.21	62.00	**76.49**	97.23
match	48.63	50.17	50.17	23.20	**76.67**	95.03
mission	71.78	71.78	71.78	78.99	**92.06**	100.00
mood	25.03	29.20	34.20	67.78	**74.28**	80.00
paper	65.47	52.79	54.54	65.23	**77.33**	99.71
post	**76.90**	59.15	65.89	34.67	48.68	96.99
pot	**58.97**	29.67	55.47	26.37	29.20	82.80
range	28.64	21.75	26.19	21.30	**50.31**	87.98
rest	35.19	39.14	42.87	25.78	**40.30**	91.08
ring	69.37	53.36	54.65	59.86	**72.23**	100.00
scene	42.67	42.67	51.00	65.94	**86.06**	90.69
side	53.75	47.03	48.27	59.62	**80.65**	93.70
soil	76.81	76.81	86.49	**96.60**	**96.60**	100.00
strain	27.40	27.40	44.44	63.66	**86.30**	94.32
test	74.55	65.29	71.66	76.21	**81.19**	91.96
Mean	**57.13**	**50.96**	**57.78**	**59.47**	**74.12**	**93.24**

Table 5.4: Upper bounds (F-Measure in %) for SemEval 2013 test dataset, obtained with different translation dictionaries: external dictionary (column **ExtDic**), dictionary based on the Multilingual Central Repository (column **MCR**), BabelNet-based dictionary (column **BabelNet**), complete GIZA++ dictionary (column **GIZA**) and pruned GIZA++ dictionary (column **GIZA10**). Last column represents results obtained by the Gold-Standard without considering multi-word translations. Bold represents best results for each word without taking the Gold-Standard into account.

5.6 Error Analysis

In this section we intend to analyse in detail the results offered by Tables 5.3 and 5.4. In particular, we want to focus on the results obtained by the ideal system using the GIZA10 dictionary. We can observe in the tables that there are some words for which other dictionaries ideally outperform the GIZA10 approach. We analyse the translation probabilities provided by this dictionary in order to look for possible explanations of this issue. Table 5.5 contains the number of translations of each word in the complete GIZA++ dictionary. After pruning the dictionary and obtaining the GIZA10 dictionary, we calculate the mean and standard deviation of the translation probabilities for each target word.

We focus on those words for which other dictionaries (external, MCR-based or BabelNet-

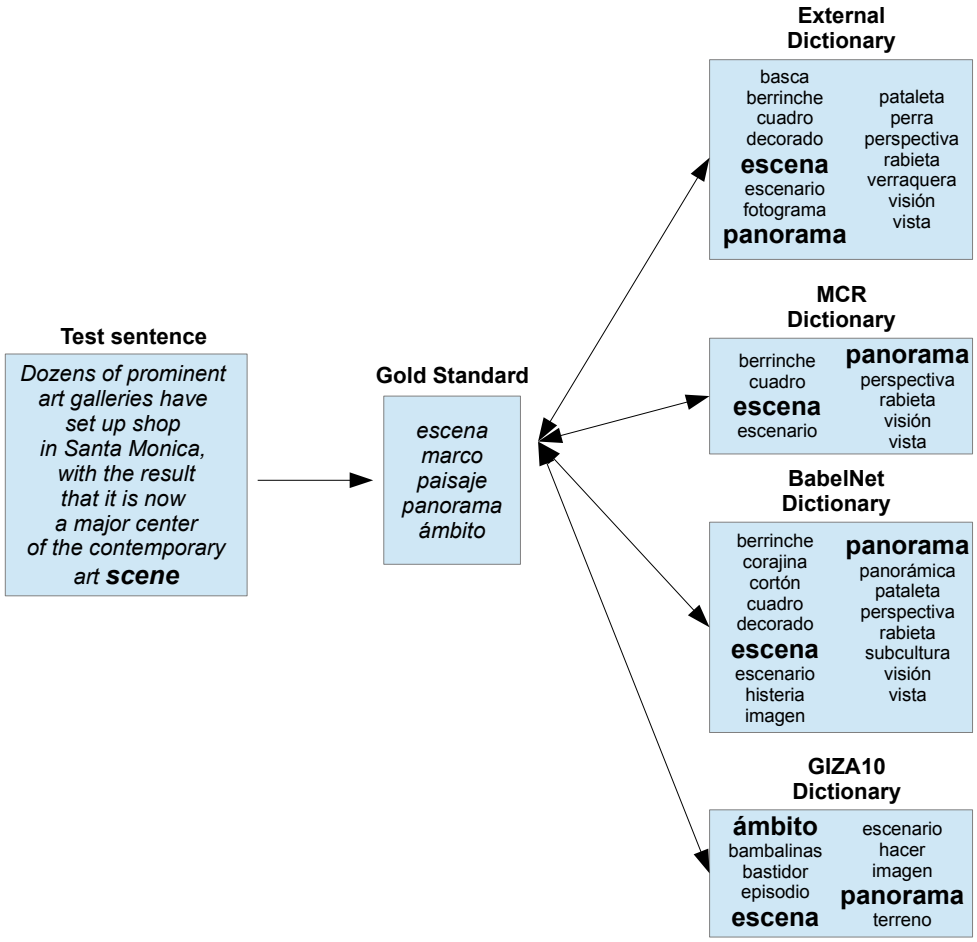


Figure 5.2: Example of the construction of the upper bounds for the considered dictionaries.

based) obtain better results for ideal systems, in both test datasets (SemEval 2010 and SemEval 2013). Those words are "match", "post", "pot", "range" and "rest". We can observe that four of those five words (excluding "post") present low mean (around 0.1) and low standard deviation (below 0.18). These facts (specially the low standard deviation) indicate that most of the translations have similar probability of occurrence, that is, the distribution adopts similar values. Hence, it is more likely that some important translations that also have a similar probability of occurrence, although slightly smaller, were discarded when the GIZA++ dictionary was pruned. Other words that present similar characteristics, such as "ring", also present worse performance in ideal systems using GIZA10, but only in one of the test datasets (in this case, SemEval 2013).

Word	Trans. (GIZA)	Mean (GIZA10)	SD (GIZA10)
coach	8	0.136	0.255
education	52	0.093	0.206
execution	30	0.104	0.301
figure	146	0.100	0.225
job	133	0.142	0.203
letter	46	0.106	0.240
match	**101**	**0.098**	**0.174**
mission	35	0.186	0.379
mood	32	0.118	0.087
paper	64	0.158	0.219
post	**72**	**0.129**	**0.219**
pot	**21**	**0.114**	**0.145**
range	**100**	**0.103**	**0.088**
rest	**87**	**0.071**	**0.162**
ring	34	0.116	0.134
scene	46	0.111	0.129
side	191	0.105	0.171
soil	10	0.126	0.295
strain	48	0.109	0.057
test	89	0.094	0.184

Table 5.5: Statistics for translations of words in the datasets. Second column contains the number of translations, third column the mean of the translation probabilities of the ten most probable translations, and fourth column the standard deviation of the same ten translations. Bold represents words for which the GIZA10 approach does not overcome the other dictionaries in neither SemEval test dataset (2010 nor 2013).

5.7 Comparison on a Particular System: CO-Graph

Once we have compared the behaviour of different dictionaries inside an ideal system, we want to consider those dictionaries inside the specific unsupervised CLWSD system described in Section 5.2. As it was stated before, the unsupervised graph construction algorithm on which the system relies depends on an initial threshold value for the p-value p. This threshold has to be determined in order to indicate the highest value of p for which the number of co-occurrences of two words is considered to be statistically significant and therefore a link is created between them.

In Chapter 4 we illustrated how the performance of our system varies when we modify the value of threshold p, from $p = 10^{-5}$ to $p = 10^{-17}$ and we selected a particular value of p for the experiments described in that chapter, based on the results shown in Figure 4.10. Those results refer to the trial dataset provided in the SemEval 2010 competition, and allowed us to

determine that the best results were obtained when the p-value was within a range between $p = 10^{-5}$ and $p = 10^{-11}$. In this chapter, as we are referring only to the English-Spanish translation, we need to focus on the results obtained in the Spanish language. However, as we can observe in the figure, those results, for the particular trial dataset, are the same for all the values of p. Hence, we need to rely on an additional criterion for selecting a threshold p. Considering the mentioned range within which the best results were obtained, we will also take into account the fact that smaller thresholds lead to smaller and more manageable graphs, in terms of resource consumptions. Because of that, we will select $p = 10^{-11}$, since it is the smallest value of p that falls into the preferred range. That threshold will be used for the experiments described in this section. By selecting a fixed threshold, we want to test the robustness of our system under the same conditions that the systems participating in the SemEval competitions. This selection of a specific value for all the cases eliminates the risk of overfitting, since known Gold-Standard data are not used for adjusting parameters.

Since we are performing a comparison between systems, it is also useful to consider a baseline for studying whether the proposed systems are able to outperform it. We take as a baseline the results obtained by a system that would return the five most frequent translations for the target word (Most Frequent Sense or MFS), according to the GIZA++ dictionary. MFS is a baseline that has been proved difficult to overcome in many CLWSD tasks, including those under analysis in this work. Moreover, these tasks use a MFS approach based on a specific corpus used to represent knowledge, and hence its performance is even better than a MFS approach based on a more generalist corpus. MFS can be extracted in an automatic way with the GIZA++ tool and has a different nature than the weights assigned by the CO-Graph system to each translation. As we proved in Chapter 4, this fact can be used for enriching the information given by the disambiguation algorithm. Hence, the combination of the weights given by the system and the probabilities given by GIZA++ may offer better results than those obtained by the original approach of our system. The intuition behind this, is based on what we stated in Section 5.6: when the values of the probabilities of translations from a target word are quite different (their standard deviation is high), CO-Graph is able to obtain a good performance, both in cases in which selecting the most frequent senses offer good results, and in cases in which the best translations do not present the highest probabilities. However, when this standard deviation of the probabilities is low, that is, when the distribution tends to be flat, CO-Graph can get lost, and hence the MFS information obtained from GIZA can be very useful.

According to this intuition, and following the scheme presented in the previous chapter, we will obtain results when combining the output of CO-Graph and the translation probabilities provided by GIZA++, by multiplying the weight w_i assigned by CO-Graph to a particular translation i, and the translation probability p_i, assigned by GIZA++ to that translation. Hence, the final score used for selecting the most appropriate translation will be $s_i = p_i w_i$.

Table 5.6 shows the performance achieved by CO-Graph, using the different considered dictionaries for both the 2010 and 2013 test datasets. It also contains the results obtained with the MFS approach, for the same datasets, and the results from the combined scheme.

Data	ExtDic	MCR	BabelNet	GIZA10	MFS	GIZA10Probs
2010	37.04	33.94	34.60	42.03	44.02	**47.41**
2013	43.87	41.35	38.95	47.06	49.75	**53.33**

Table 5.6: Results (F-Measure in %) obtained over 2010 and 2013 SemEval test datasets, for the out-of-five evaluation. Columns 2 to 5 contain the results achieved by the CO-Graph system when using the different bilingual dictionaries (external, MCR-based, BabelNet-based and GIZA++ pruned to ten translations per word). Column 6 represents the results obtained by the MFS (Most Frequent Sense) approach. Last column shows results obtained by the combination of the system output with GIZA++ probabilities.

The results clearly show, on one hand, that the test dataset for the 2013 competition allows the system to obtain a higher performance. This is basically due to the use of the same words as in the 2010 competition, but modifying the contexts for evaluation. As we can observe, all approaches improve their performance from 2010 to 2013. On the other hand, we can observe that, as we expected, the use of the GIZA10 dictionary, allows the system to improve the results, when compared to those obtained with the other three dictionaries. We observe that the F-Measure achieved by the system using a particular dictionary is directly proportional to the average number of translations for each word in the dictionary, in a similar way to what happened with the ideal systems. As we stated above, we performed different tests regarding the pruning value of the GIZA++ dictionary, observing that when more than 10 words were used as maximum number of translations for each word, the performance of the system decreased drastically. For example, we have tested the CO-Graph system built with a dictionary obtained through GIZA++ but pruned with up to 20 translations per word, maintaining the same configuration parameters (disambiguation algorithm, value for the threshold when building the graph), and the achieved F-Measure for the 2010 test dataset is 32.20%, that is, around 10 points below the results obtained with the GIZA10 dictionary, and 15 points below the GIZA10Probs configuration. Hence, the key point of pruning the GIZA++ dictionary is to find a large enough maximum number of translations (coverage of the problem) that does not introduce too much noise into the system. Table 5.6 shows that the value of 10 translations per word offers good results. Since we select those translations with highest probability of occurrence, the overall performance of the system is better than that achieved when using the MCR-based dictionary for instance, a dictionary that uses a similar (average) number of translations for the target words in the datasets (see Table 5.2). Still, the Most Frequent Sense technique outperforms any of the proposed approaches. This fact indicates that when more than five translations are considered, the system does not effectively choose the most suitable ones. We can also observe that, for both datasets, the last approach (GIZA10Probs, which uses GIZA10 and combines the output with translation probabilities) gets better results than the MFS approach. The performance of the system increases about 3.5 points for the 2010 and 2013 datasets. Moreover, the improvement over the system that uses the pruned GIZA dictionary (GIZA10) is more than 5 points in the 2010

dataset and more than 6 points in the 2013 dataset. A two-tailed paired t-test for statistical significance testing has been performed over the results in the table. According to this test, the results obtained by this last approach are significantly better than those obtained by the system using only any of the bilingual dictionaries. Also, in the 2010 dataset, the differences between the GIZA10Probs and the MFS approach are statistically significant, whereas in the 2013 dataset, although the results are also better, the significance is not achieved.

A last experiment has been conducted regarding the words used as nodes of the co-occurrence graph. As we have stated along the previous and present chapters, we have used only nouns for building the co-occurrence graph of the CO-Graph system, and for extracting the context of the target word in each test sentence. We are interested in testing whether the selection of other important category of words, in this case verbs, could improve the overall performance of the system. For this purpose, a new co-occurrence graph that also considered verbs was built, and the disambiguation process was repeated for all the test instances, extracting also verbs from the context. Table 5.7 shows the comparative between results obtained by the system using only the best dictionary (GIZA10) and by the system that combines those results with translation probabilities (GIZA10Probs), both using only nouns and using nouns and verbs for building the graph and extracting the context.

	Words	**GIZA10**	**GIZA10Probs**
SemEval 2010	Nouns	**42.03**	**47.41**
	Nouns+Verbs	34.70	45.00
SemEval 2013	Nouns	**47.06**	**53.33**
	Nouns+Verbs	39.58	51.33

Table 5.7: Comparative between results obtained by the best performing configurations of the system (GIZA10 and GIZA10Probs), using only nouns for building the graph and extracting the context, and using nouns and verbs for these processes. Results (F-Measure in %) for the 2010 and 2013 test datasets.

As we can observe, the inclusion of verbs in the construction of the graph does not improve the results. Including new words in the graph may lead to bigger, more difficult to handle graphs, and hence to more difficulties in the disambiguation process. Also, it is important to indicate that most of the target words in the test instances can be translated as nouns. Therefore, the increase of coverage that could be achieved by including verbs in the translations may not compensate the probable loss of precision due to the need of dealing with bigger graphs.

5.8 Comparative

In this section we will compare the results obtained by the best configuration of our system (GIZA++ dictionary pruned to 10 translations per word, and combined with translation probabilities), with other systems participating in the 2010 and 2013 SemEval CLWSD competitions. Some of these results have already been shown in Section 4.7.6 of Chapter 4, however, in these tables results for every unsupervised system participating in the English-Spanish part of the 2010 and 2013 CLWSD tasks are shown. Table 5.8 shows results for the 2010 competition, whereas Table 5.9 shows results for the 2013 competition. Results obtained by the best participating system (even if supervised) are also shown, as well as the baselines proposed in the competitions.

System	Task 3 SemEval 2010
Best	43.12
CO-Graph	**47.41**
T3-COLEUR	35.65
UHD-1	34.95
UHD-2	34.22
Baseline	48.41

Table 5.8: Comparison of the F-Measure (%) achieved by the unsupervised systems participating in the English-Spanish part from task 3 of SemEval 2010, and by the best configuration of our system (row **CO-Graph**). The best participating system (even if supervised) is shown in row **Best**, while the baseline proposed by the organizers is shown at the bottom of the table, in row **Baseline**.

System	Task 10 SemEval 2013
Best	61.69
CO-Graph	**53.33**
LIMSI	49.01
XLING snt	44.83
XLING merged	43.76
XLING tnt	39.52
NRC-SMT adapt2	41.65
NRC-SMT basic	37.98
Baseline	53.07

Table 5.9: Comparison of the F-Measure (%) achieved by the unsupervised systems participating in English-Spanish part from task 10 of SemEval 2013, and by the best configuration of our system (row **CO-Graph**). The best participating system (even if supervised) is shown in row **Best**, while the baseline proposed by the organizers is shown at the bottom of the table, in row **Baseline**.

As we can observe, in both cases CO-Graph outperforms the results obtained by other

unsupervised systems. More particularly, the unsupervised systems in the 2010 task were the T3-COLEUR system, based on probability tables, and the UHD system, also based on co-occurrence graphs, but with different techniques for extracting the knowledge from the graph to perform the disambiguation. In the 2010 competition, we can also see that the best participating system (supervised) is also outperformed by CO-Graph. However, the baseline proposed by the organizers is still the best "system" in the task. We consider this baseline to be an unrealistic approach to the problem, since not even supervised techniques are able to outperform it. In 2013, the unsupervised participants were the vector-based LIMSI system, the XLING system, using topic modelling techniques, and the NRC system, based on a statistical machine translation approach. Regarding this dataset, we observe that the best (supervised) system is better than CO-Graph. In this case, the proposed baseline is outperformed by our system, but not by any of the unsupervised systems that participated in the competition. More detailed definitions of the participating systems can be found in Chapter 2.

5.9 Conclusions

We have analysed the effect of the translation dictionary in the performance of a Cross-Lingual Word Sense Disambiguation system. The results obtained within an ideal framework indicate that when the dictionary is generated in a statistical automatic way from a corpus large enough to represent the characteristics of a language, the potential results for a disambiguation task are better. The best ideal results are achieved when considering all the possible translations obtained. However, this induces too much noise. Accordingly, the number of potential translations for each word has been pruned to the ten most probable ones for building the GIZA10 dictionary. For some target words, the other dictionaries are able to outperform the results obtained by an ideal GIZA-based dictionary. This fact can be due to the nature of the set of translations that GIZA extracts for each word: when too many translations are extracted, and their probabilities are similar, the coverage that can be achieved by a dictionary containing ten translations per word can be compromised. Nevertheless, the GIZA10 dictionary has been shown to be the best dictionary in ideal conditions. This selection has been confirmed using a particular CLWSD system. CO-Graph has been tested over the four different dictionaries, and the results have been compared to those obtained by a Most Frequent Sense (MFS) baseline. In this case, the GIZA10 dictionary has also proven to be the best choice among the analysed dictionaries for solving the CLWSD tasks. However, the MFS approach still outperforms its results. Considering this fact, and the unsupervised nature of the MFS approach, a last approach has been built, using outputs from both CO-Graph and the MFS approach. The results obtained by this approach outperform the MFS reference baseline, and the other unsupervised systems participating in the 2010 and 2013 CLWSD competitions from SemEval. The main conclusion we can extract from this finding is that statistical information related to the possible translations of the target

words, is a key knowledge for systems performing CLWSD, as we already introduced in Section 4.8 of Chapter 4. Accordingly, this way of selecting the candidate translations can be considered as one of the best options for unsupervised CLWSD systems.

6

WORD SENSE DISAMBIGUATION IN THE BIOMEDICAL DOMAIN

The chief virtue that language can have is clearness, and nothing detracts from it so much as the use of unfamiliar words.

Hippocrates

Contents

6.1 Introduction

In the introduction of this thesis (Chapter 1) we already mentioned our intention of applying the work and techniques developed on the field of Word Sense Disambiguation to a more specific field, in this case the biomedical domain. As we stated in that introduction, the amount of information available in this field of study has lead to the development of many NLP systems which eventually rely on disambiguation techniques to improve their behaviour. Moreover, there exist many different types of lexical ambiguity in biomedical documents, which represents an additional challenge when performing WSD in this domain (Stevenson and Guo, 2010a): words and phrases with more than one possible meaning, abbreviations with more than one possible expansion, or names of genes which may also contain ambiguity when standard naming conventions are not followed (the names of more than one thousand gene terms are standard English words (Sehgal et al., 2004)). The use of biomedical concepts, in addition to plain text, when working with medical documents, can be seen as another challenge, since the process of transforming plain text into biomedical concepts is an additional step not considered when working with more general texts, that is, not belonging to a specific domain (Savova et al., 2008). In this chapter, we present a modified version of the disambiguation technique based on co-occurrence graphs presented in Chapter 3 for addressing the Word Sense Disambiguation (WSD) problem in the biomedical domain. The expected contributions of this chapter are to evolve the graph-based approach for its use in the biomedical domain and, by evaluating it using datasets containing a range of ambiguities, demonstrate that it outperforms alternative approaches that do not make use of external knowledge sources. As we stated in Chapter 4, when we tackle Cross-Lingual WSD tasks, raw text annotated with POS tags is the most useful form to represent the information that populates the co-occurrence graph, since the ambiguity introduced by a given word in the source language (in our case English) can be represented by the set of potential translations in the target language. However, when it comes to biomedical WSD, our task is to determine the most suitable sense of a target ambiguous biomedical term, between all the biomedical concepts to which the ambiguous term could refer (meanings of a word or phrase, expansions of an abbreviation, etc.). For solving this issue, we need to find a way to represent or name the collection of possible senses an ambiguous biomedical term may refer to. The most widely used knowledge source in the biomedical domain, as stated in Chapter 2, is the Unified Medical Language System (UMLS) Metathesaurus (Humphreys et al., 1998), a biomedical database which assigns a Concept Unique Identifier (CUI) to each biomedical concept in an unequivocal way. This way, in this domain the ambiguity introduced by a term is represented by the set of possible CUIs related to its textual form. Thus we need a way to transform our text documents into these CUIs representing biomedical concepts, as we will explain later on.

In this chapter we describe the application of our CO-Graph technique for performing WSD in the biomedical domain in a monolingual context, while a multilingual approach to the

problem is presented in the following one, Chapter 7.

The rest of the chapter is organised as follows: Section 6.2 describes the proposed system, detailing the different steps involved in the disambiguation process. Evaluation is carried out using two datasets (see Section 6.3) with the results described in Section 6.4. Finally, conclusions are found in Section 6.5.

6.2 System Description

For the creation of the co-occurrence graph, we will use the same theoretical background described in Chapter 3, based on our hypothesis that documents in a corpus are consistent, and hence that concepts in document present a strong tendency to be related. The statistical analysis applied to identify those concepts in documents that fulfill our hypothesis will be the same, although the nature of the concepts that will eventually populate the co-occurrence graph will slightly change. These changes will be part of a step prior to the construction of the graph, called annotation phase. We have shown in Chapter 4 that the technique offers successful results when applied to general WSD tasks such as Cross-Lingual WSD. This fact suggests that a similar approach could also lead to competitive results in domain-specific WSD.

In this section, the annotation phase, as well as all the steps involved in the disambiguation, are detailed.

6.2.1 Annotation

The first step in the creation of the co-occurrence graph is to annotate the biomedical concepts that appear in the documents. These concepts will eventually become the nodes of the co-occurrence graph which forms the knowledge base used by our system. As we introduced in Section 6.1, in this case we will not need the raw textual information for populating the graph, but the CUIs representing the concepts to which a biomedical term written in its textual form may refer. The annotation step consists in transforming the plain text that can be found in the medical documents, into CUIs that represent equivalent medical concepts. This step could be carried out by manual annotation, although in this case we perform it automatically, through the MetaMap program (Aronson, 2001), which allows us to split the text inside a document into phrases, and map each of those phrases onto a set of UMLS CUIs. Once we have transformed each of the text documents in a corpus into a set of CUIs, we are able to calculate the co-occurrences between those CUIs for building the graph. The MetaMap program offers the possibility of using a disambiguation server which helps

the user to select a candidate for each phrase in the text. We make use of this server when annotating the documents that will be used for building the document graphs. If it was not used, each time an ambiguous medical concept appeared in a document it would be replaced by all its possible senses (CUIs) and consequently the co-occurrences would not provide useful information for performing the disambiguation. However, as the configuration of the disambiguation server can be set when running the program, we have selected unsupervised methods for this initial disambiguation in all the experiments conducted. This way, we assure that the unsupervised nature of our system is maintained through all the process. A baseline containing the results obtained by the disambiguation server considered in our experiments will be reported in subsequent sections. As we will see, the quality of this disambiguation is far from the results achieved by our system. We maintain the default values for the rest of the configuration parameters when running the MetaMap program.

Figure 6.1 illustrates the annotation process of a text document using MetaMap: The program generates the mappings for each of the phrases extracted from the text that can be matched against CUIs from UMLS. After that, we extract all the proposed CUIs and represent each of the documents in the corpus as a set of CUIs. Those final documents will then be used for calculating the co-occurrences between CUIs and building the co-occurrence graph.

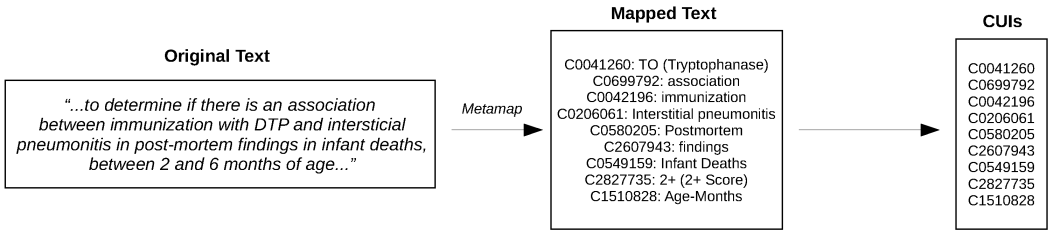


Figure 6.1: Annotation of a biomedical document.

6.2.2 Graph Construction

As explained in Chapters 4 and 5, the construction of the co-occurrence graph follows the statistical procedure detailed in Chapter 3. This way, although in this case the elements that will populate the co-occurrence graph are CUIs from UMLS instead of words from raw text, the calculation of the statistical significance of a co-occurrence between two concepts will be conducted in the same way. The final result of this process will be a co-occurrence graph in which nodes are CUIs from UMLS and weights of the edges represent the significance of the co-occurrence of two CUIs.

6.2.3 Disambiguation

Once that we have built our co-occurrence graph, we need to define a disambiguation algorithm, as we did in Chapter 4. This algorithm will allow us to determine the most suitable sense (CUI) of an ambiguous biomedical term given its context, among all the possible senses provided by a dictionary. In other general tasks such as CLWSD, the selection or construction of this dictionary is a key point for assuring the good performance of a system, as we proved in Chapter 5. However, in this domain the dictionary that contains the possible senses of every target word is automatically provided within the test dataset.

The disambiguation algorithm that we have selected for performing this last step is the Personalized PageRank algorithm, initially introduced by Haveliwala (2002), and based on the PageRank algorithm (Brin and Page, 1998). The algorithm, already introduced in Chapter 4, is based on the following main formula:

$$P = cMP + (1 - c)v, \tag{6.1}$$

where P is the vector that contains the PageRank values for each node, c is a constant called "damping factor" usually set to 0.85, M is the matrix containing the values of the out-degrees of the nodes and v is a $N \times 1$ stochastic vector, being N the number of nodes in the graph. We will maintain the default value of the damping factor, this is, $c = 0.85$.

As we did in the "PageRank with Priors" approach, in Section 4.6.2 of Chapter 4, some nodes of the graph are initially powered up using vector v. In this case, we will use those nodes that represent CUIs appearing in the context of the target concept we want to disambiguate. Hence, before performing the disambiguation step, we need to convert the plain text of each test instance onto the set of CUIs that represent all the medical concepts that can be found in the text, also using the MetaMap program. When a term in the text is ambiguous, MetaMap assigns all the possible CUIs that may correspond to it. When it comes to a target concept, this set of possible CUIs becomes the ambiguity that our system is trying to solve. That is, the disambiguation service provided by MetaMap is not used in this step, since it would give us a priori information about the possible senses of the concepts in the context of the target word. The rest of the configuration parameters are also set to their default values.

Once that we have all the CUIs that belong to the context of the target concept, we build v as a $N \times 1$ vector whose values will be $v_i = \frac{1}{C}$ if node i represents a CUI of the context, and 0 otherwise, being C the total number of CUIs found in the context of the target concept. After performing the Personalized PageRank algorithm, we will select the node with highest rank, among those representing possible senses of the target concept.

Figure 6.2 shows an example of successful disambiguation, by illustrating the behaviour of the Personalized PageRank (PPR) disambiguation on our co-occurrence graph, and

comparing it with the result obtained by running PPR over a graph directly built from the UMLS database. In this UMLS graph, two nodes are linked together if a relation between them can be found in the UMLS database. The example is divided in two parts: the top part of the figure presents a test instance which contains the target word "culture", to be disambiguated. A look-up to the dictionary tells us that the two different senses (CUIs) of "culture" between which our system should discriminate are "C0430400", referred to a microbial culture (laboratory process), and "C0010453", referred to a culture from an anthropological point of view. Then, we obtain all the CUIs that represent concepts from the context of the test instance by applying MetaMap to the text.

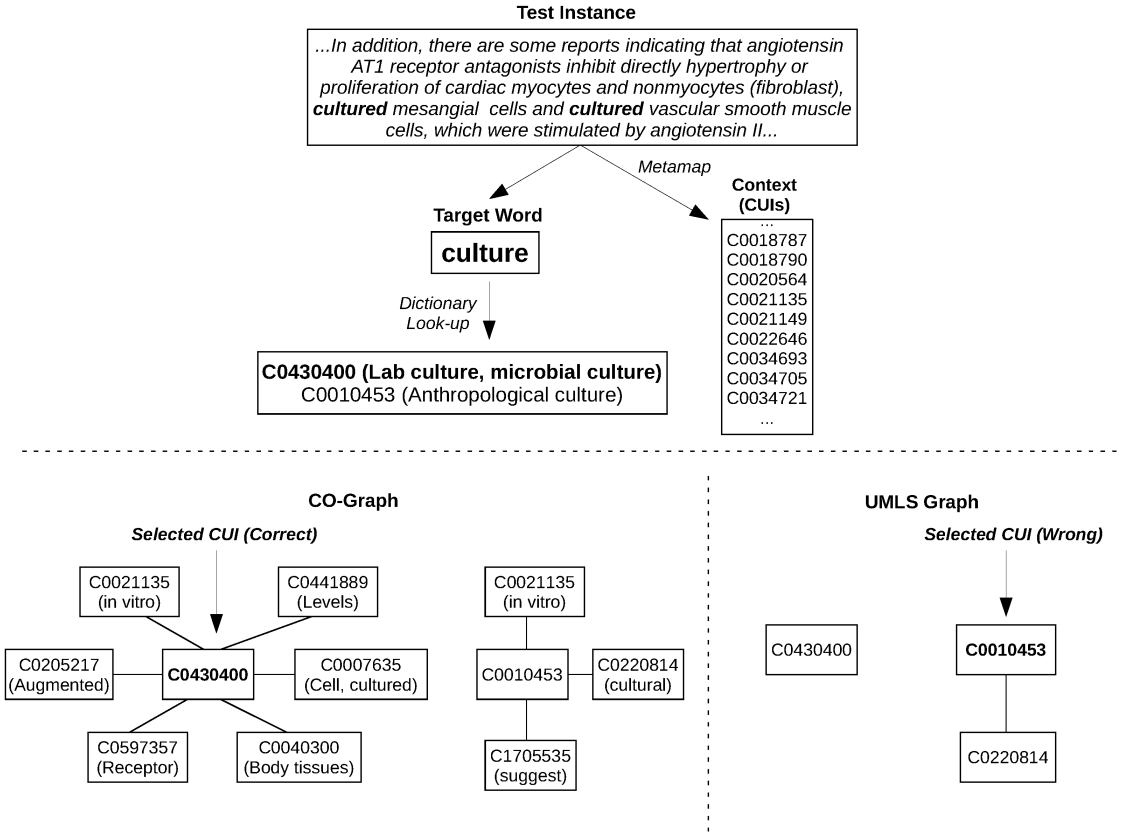


Figure 6.2: Example of disambiguation. Extraction of the target and context CUIs (top part) and comparison between the disambiguation algorithm over the co-occurrence graph and the UMLS fixed graph (bottom part).

The second part of the figure (bottom part) illustrates the differences of applying the disambiguation process using our co-occurrence graph, or the UMLS graph. In our co-occurrence graph the correct sense of "culture" ("C0430400") is much more related to context CUIs than the other sense ("C0010453"). Hence, the disambiguation algorithm selects this more connected sense to be the most appropriate for this test instance. However, when using the UMLS graph we can observe that both senses are poorly connected to the CUIs in the context (which will result in a higher randomness when selecting a sense). In fact, the wrong sense is connected to one CUI in the context, while the correct sense is not connected to any of the CUIs in the context. Because of that, the disambiguation algorithm mistakenly selects

the CUI “C0010453” to be the most appropriate for this test instance.

6.3 Datasets

This section describes the datasets used to evaluate our system.

6.3.1 Acronym Corpus

The Acronym corpus (Stevenson et al., 2009) contains 55,655 abstracts downloaded from Medline. Each of these abstracts contains an ambiguous acronym from a set of 21 originally developed by Liu et al. (2001) and widely used in previous research. These acronyms each consist of at least 3 letters and are associated with between 2 and 5 extended forms (which are considered as senses). The dictionary for the target concepts is then created using the CUIs that correspond to their possible extended forms. A small subset of the corpus is split into three different test datasets, containing 100 instances, 200 instances and 300 instances per ambiguous acronym. We will refer to those datasets as “A100”, “A200” and “A300”, respectively. However, not all the 21 acronyms are present in every dataset, since some of them were removed from the test datasets due to an insufficient number of instances in the main corpus. Also, some acronyms such as “ACE”, “ASP” and “CSF” were also removed from the initial datasets, in order to reduce their imbalance, since most of their test instances belonged to the same extended form. As a result, the A100, A200 and A300 datasets contain 18, 16 and 14 different ambiguous acronyms respectively. The final dataset obtained after this pre-processing is the same used by other state-of-the-art techniques to which we compare our system.

Data acquisition: Since the corpus was initially created for a supervised system, all the abstracts are annotated with the extended form that corresponds to the acronym found in the text. Since our system is unsupervised, it does not need these annotations, however, we need to acquire data to build our co-occurrence graph. These data will be represented by the abstracts from the original corpus that are not included in any of the three test datasets. Hence, our co-occurrence graph will be created from a set of 50,143 abstracts, which will be previously mapped onto CUIs from the UMLS database, as explained in Section 6.2.1.

6.3.2 NLM Corpus

The second corpus we will use to evaluate the performance of our system is the NLM-WSD corpus (Weeber et al., 2001). In contrast to the Acronym corpus, this corpus is composed of general ambiguous terms. It contains 50 terms with 100 instances per term. These instances are also abstracts downloaded from Medline, and manually annotated with the CUI that represents the correct sense for the target term in each instance. However, during the creation of the corpus, annotators could select to mark as "None" those instances for which none of the possible senses applied. We have removed those instances, so the final test dataset, which will be referred to as "NLM", contains 3,983 instances and 49 terms (since all the instances were marked as "None" for the term "association"). As with the Acronym corpus, the same pre-processing is applied to the state-of-the-art techniques against which our system is compared.

Data acquisition: In this case, given that the NLM-WSD corpus is a test dataset itself, we do not have a set of documents to build the co-occurrence graph. For this purpose, we downloaded our own set of abstracts from Medline, using the Entrez interface (Sayers, 2013). We performed a search for each ambiguous term of the test dataset, restricting the results to 1,000 abstracts per term. In order to avoid downloading abstracts that could appear in the test dataset, we have only downloaded abstracts from year 2014. For maintaining the unsupervised nature of our technique, we do not specify in any way the sense of the ambiguous term for performing the search, so in the downloaded abstracts any possible sense of the target term can be found. The total number of abstracts in this set is 35,282. Although we downloaded 1,000 possible abstracts for each of the 50 ambiguous terms in the dataset, there are abstracts containing more than one term, and hence the reduction of the number of documents in the final corpus [i].

6.3.3 Dataset Properties

Table 6.1 resumes the characteristics of the datasets used for evaluation. We can observe that for datasets "A200" and "A300" there is one abstract missing (given the number of ambiguous terms, and instances per term, they should have 3,200 and 4,200 instances respectively). This missing abstract was no longer available for download from Medline. The average number of possible senses is higher in the Acronym corpus than in the NLM corpus, although the total number of ambiguous terms is quite higher in this last corpus.

[i] Corpus available at nlp.uned.es/∼aduque/NLM_related_public.zip

	A100	A200	A300	NLM
Instances	1,800	3,199	4,199	3,983
Amb. terms	18	16	14	49
Min/Max # senses	2 / 4	2 / 4	2 / 4	2 / 5
Avg # senses	2.61	2.5	2.57	2.24

Table 6.1: Statistics for the different test datasets: number of instances, number of ambiguous terms (or acronyms), minimum and maximum number of senses for a term and average number of senses per term.

6.4 Evaluation

This section presents the results obtained by the approach described here and compares them with other state-of-the-art systems. Also, an exhaustive analysis of the parameters of the system is also performed, in order to study how the results vary depending on their values.

6.4.1 System Results and Comparison

As we stated in previous sections, a co-occurrence graph was built for each of the evaluation corpus: the Acronym corpus (whose graph was used for evaluating the three test datasets, “A100", “A200" and “A300") and the NLM-WSD corpus. The performance metric used to evaluate the system performance in all experiments is accuracy: number of correctly disambiguated instances divided by the total number of instances in the test dataset, expressed in %. Table 6.2 shows the accuracy achieved by our system in each of the test datasets. In order to analyse the impact of the selected co-occurrence graph when evaluating the system, we have also included the results obtained by cross-testing our graphs, this is, using the graph created with abstracts from the Acronym corpus for evaluating the “NLM“ dataset, and vice versa. Finally, a joint graph was created combining the 50,143 abstracts of the “non-test" Acronym corpus and the 35,282 abstracts of the acquired “NLM-WSD related” corpus. The results of applying this joint graph to all the test datasets are also shown in the table.

Results show that the graph created with abstracts from the Acronym corpus produces similar results on the three acronym test datasets. Regarding the cross-testing experiment, the results obtained using the Acronym-based graph over the NLM dataset are similar to those obtained by using the NLM-based graph over the NLM dataset. However, the NLM-based does not perform as well in the cross-testing scenario, i.e. when applied to the Acronym datasets.

	Datasets			
	A100	**A200**	**A300**	**NLM**
Acronym Graph	82.11	79.87	**82.64**	74.24
NLM Graph	61.83	59.59	58.83	75.45
Joint Graph	**82.78**	**80.06**	82.57	**78.36**

Table 6.2: Results (accuracy in %) for the co-occurrence graph-based system, for each of the graphs (Acronym corpus, NLM-related acquired corpus and joint graph), in each of the different test datasets. Bold highlights the best result obtained for each of the test datasets.

This may be due to a greater specificity of the Acronym corpus, in which the different CUIs among which the disambiguation algorithm has to choose (representing extended forms of the acronyms), correspond to more specific concepts. On the other hand, terms in the NLM-WSD corpus are much more general. Hence, it is possible that some of the target CUIs of the Acronym corpus do not even appear in the graph created from NLM-related abstracts. Also, it is likely that any graph created from a large enough set of abstracts (such as the one created with acronym-based abstracts) contains enough information about CUIs representing the general concepts of the "NLM" dataset to perform a good disambiguation. Finally, we can observe that results obtained with the joint graph improve those obtained with simpler graphs for all but one of the datasets. This suggests that the combined information that can be found inside the joint graph is useful to better represent the connections between concepts and hence help to improve the overall disambiguation.

6.4.2 Comparison with Previous Approaches

Table 6.3 shows a comparison between the results obtained with our co-occurrence graph-based system ("CO-Graph" in the table) as well as other knowledge-based and unsupervised systems that present results for the same datasets. The "NLM" dataset is more commonly used for evaluation than the Acronym datasets in the literature.

The first two rows of the table show results obtained using two different baselines: in the first row, we have the "Most Frequent Sense" (MFS) approach, which can be considered as a supervised baseline, and represents the accuracy achieved by a system that classifies every instance as belonging to the most common CUI for its ambiguous term. As we can observe, the MFS value for the NLM dataset is high demonstrating that it is imbalanced (i.e. for many of the ambiguous terms most of the instances belong to the same CUI). Also, we show results obtained by running the MetaMap program against the test dataset, and making use of the disambiguation server under the same conditions we used for annotating the documents when building the co-occurrence graph, as explained in Section 6.2.1. As we can observe, the results for the NLM dataset are quite low in comparison with the accuracy achieved by

	Datasets			
	A100	**A200**	**A300**	**NLM**
MFS	69.00	69.10	68.70	84.71
MetaMap	—	—	—	49.13
PPR+UMLS	56.33	56.99	58.02	68.10
AEC	—	—	—	68.36
JDI	—	—	—	74.75*
MRD	0.57	0.61	0.61	63.89
2MRD	**88.00**	**90.00**	**89.00**	55.00
CO-Graph	82.78	80.06	82.57	**78.36**

Table 6.3: Comparative of results (accuracy in %) for state-of-the-art systems (see text) and the system reported here for our co-occurrence graph-based system (CO-GRAPH) in each of the different test datasets. Bold highlights the best unsupervised results obtained for each of the test datasets.

our system. Since the MetaMap program does not offer any disambiguation for acronyms, this second baseline does not offer results for the A100, A200 and A300 datasets.

Results from our system are compared against different WSD systems, mentioned in Chapter 2: The **PPR+UMLS** system (Agirre et al., 2010) uses a graph-based similar approach, which makes use of a fixed graph built from the UMLS database, as described in the example shown in Figure 6.2. Although in the original work it is only applied to the "NLM" dataset, we have also reproduced this technique for testing the Acronym datasets, in order to obtain a better comparison. The **AEC** (Automatic Extracted Corpus) system (Jimeno-Yepes and Aronson, 2010) is a semi-supervised approach that automatically downloads and annotates abstracts for training a machine learning system. The **JDI** (Journal Descriptor Indexing) method (Humphrey et al., 2006) makes use of semantic type vectors that represent each possible sense of an ambiguous term and computes their distance to a vector representing the test instance. Although it obtains good results for the NLM corpus, it only takes into account those senses belonging to different semantic types, hence many instances of the NLM corpus were removed in this experiment. That is the reason why results obtained by this system are marked with an asterisk in the table. Finally, the **MRD** and **2MRD** techniques are applied in (McInnes, 2008) and (Jimeno-Yepes et al., 2011) over the NLM corpus, while results achieved by the 2MRD technique over the Acronym datasets are presented in (McInnes et al., 2011).

As we can observe in Table 6.3, our system outperforms all the state-of-the-art knowledge-based and unsupervised methods when applied to the NLM dataset, and even semi-supervised ones. Regarding the improvements obtained by our method with respect to the one that uses relations from the whole UMLS graph (PPR+UMLS), which can be considered the most similar approach to ours, we consider that contextual information obtained from actual abstracts in the process of building the graph is able to better represent knowledge that may eventually lead to correctly disambiguate a term inside a different abstract. Relations from

the UMLS graph can be useful, but they do not necessarily imply that two related terms are likely to co-occur in the same document. The second-order vector technique (2MRD) outperforms our system in the Acronym corpus. However, while this technique makes use of additional information from UMLS (extended definitions of the possible senses), the main contribution of our method is that our disambiguation phase is completely based on the co-occurrence graph created from the abstracts, so it does not need additional information from the UMLS database.

6.4.3 Parameter Analysis

In this section we explore the effect of varying the two parameters used by the approach described here. The joint graph (built with abstracts from both the Acronym and the NLM-related corpus) is used for the experiments described here.

The first parameter is the threshold for the p-value p (see Chapter 3). This threshold, denoted by p_0, establishes the highest accepted value for p in order to consider a co-occurrence to be statistically significant, and hence create a link in the graph between the two co-occurring CUIs. Figure 6.3 illustrates the behaviour of our system, in terms of accuracy for each test dataset, when we vary p_0, decreasing its value from $p_0 = 10^{-2}$ to $p_0 = 10^{-11}$. We have chosen a maximum value of as 0.01, since experiments in which greater thresholds were used showed that the resulting graphs are unmanageably large and performance quickly decreases.

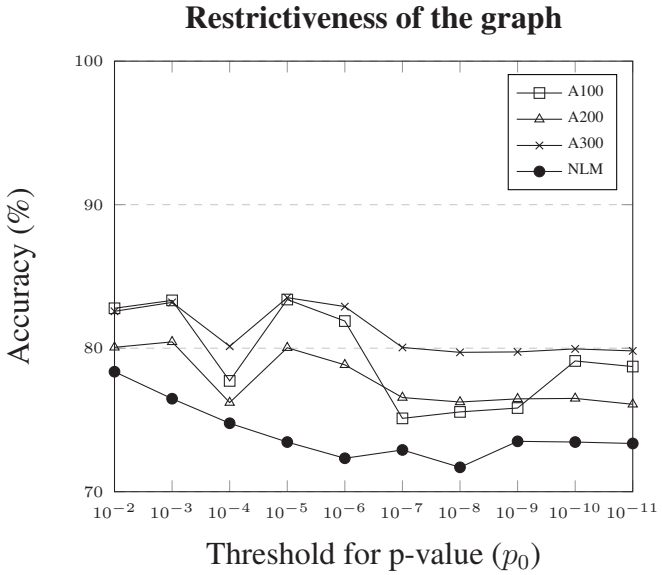


Figure 6.3: Evolution of the accuracy (%) as the specified threshold for the p-value decreases (the restrictiveness of the graph increases).

As we decrease the threshold, it is more difficult for a pair of CUIs to present a statistically significant p-value, and hence the graph becomes more restrictive, reducing the number of edges. The best results are obtained for the least restrictive graphs, while accuracy usually decreases as we decrease p_0. This is due to the removal of important edges representing relations between concepts, as we increase the restrictiveness of the graph.

Figure 6.4 represents the behaviour of the system depending on the number of abstracts used for building the co-occurrence graph. The complete set of abstracts used for building the joint graph was randomized, and gradually larger subsets of those abstracts were used to build the graphs. As we increase the number of abstracts, each subset contains all the abstracts of the previous one.

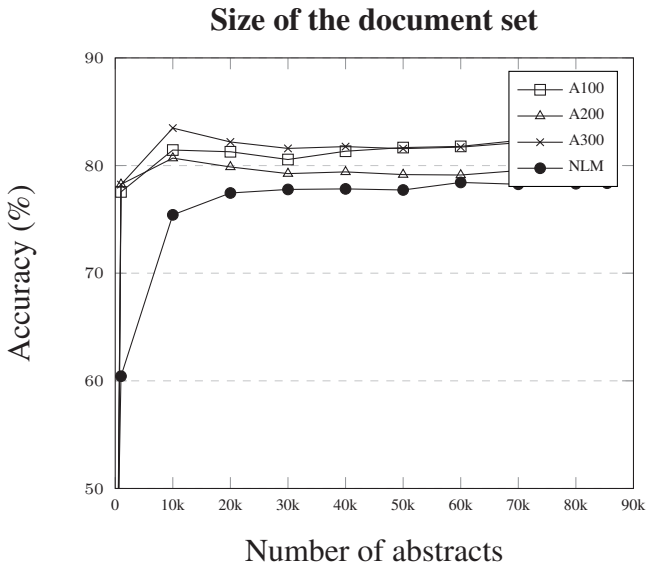


Figure 6.4: Evolution of the accuracy (%) as the number of abstracts used for building the co-occurrence graph increases.

The overall accuracy increases with the number of abstracts used to build the graph, although performance for each method quickly reaches a plateau. Results rapidly converge to an accuracy of more than 80% in the A100, A200 and A300 datasets, and around 77% in the NLM dataset. Fast convergence of the algorithm is a useful feature when resources are limited.

6.5 Conclusions

In this chapter we have described the application of our technique based on co-occurrence graphs for performing WSD in the biomedical domain. The knowledge base on which the system relies is automatically created in an unsupervised way from a set of abstracts downloaded from the Medline database and automatically mapped onto medical concepts. Unlike

other state-of-the-art techniques, external resources are not used for the disambiguation step. Evaluation on two widely used test datasets shows that the reported method obtains consistent results that outperform most of the knowledge-based systems addressing the same problem. Through these experiments and their correspondent evaluation, we have proved the validity of the CO-Graph technique for performing WSD in the biomedical domain, and its robustness when working with different types of concepts for building the co-occurrence graph, such as CUIs, instead of words as we did in previous chapters. Further experiments suggest that the convergence of the method is fast regarding the number of abstracts used for building the graph. In addition, better results are obtained with less restrictive graphs, since they incorporate to the co-occurrence graph the most useful information about relations between concepts for performing the disambiguation.

7

MULTILINGUALITY FOR BIOMEDICAL WORD SENSE DISAMBIGUATION

We die. That may be the meaning of life. But we do language. That may be the measure of our lives.

Toni Morrison

Contents

7.1 Introduction

Widely explored in the NLP literature, multilinguality has been proven to be a really useful source of information when it comes to NLP tasks (Faruqui, 2014; Fernandez-Ordonez et al., 2012; Huang et al., 2013). The use of multilingual data could palliate the lack of information present in some fields of the biomedical domain, as we introduced in Chapter 2. Hence, one of the initial hypotheses of this chapter considers that significant improvements can be achieved in NLP tasks in the biomedical domain by adding multilingual information to a knowledge-based system. In particular, and following the research line we have studied throughout this thesis, we will focus on the Word Sense Disambiguation task and the benefits it can extract from multilinguality. As we stated in Chapter 6, one of the main challenges in the field of biomedical WSD is the existence of different sources of ambiguity (words and phrases with different meanings, acronyms with different expansions, etc.). In the previous chapter we avoided the use of plain text since we tackled biomedical WSD from a monolingual perspective, in which concepts could be transformed into identifiers (CUIs from the UMLS biomedical database). However, the addition of multilingual information forces us to consider not only concepts written in English, which can be transformed into CUIs, but also concepts written in different languages, hence the use of text directly extracted from the corpus used as source of information will be mandatory for representing the information provided by those other languages.

It is difficult to find works in the literature that apply multilinguality to the WSD task in the biomedical domain, probably due to the lack of bilingual corpora providing enough useful information for disambiguation, that is, a wide enough collection of documents containing ambiguous terms, and with a balanced number of occurrences for each possible sense of such terms.

The main objective of this chapter is the application of multilingual techniques to an adapted version of our unsupervised graph-based approach already described in previous chapters of this thesis (CO-Graph), for performing WSD in the biomedical domain. This way, we intend to analyse the improvements that can be achieved with the addition of multilingual data to our knowledge base, by evaluating this multilingual system on widely known datasets containing a range of ambiguities. We perform a thorough analysis of the conditions under which the proposed approach becomes a useful and powerful tool to solve the WSD problem. We also explore ways of dealing with the lack of available bilingual corpora, as well as different languages and their contributions to possible improvements.

The rest of the chapter is organized as follows: Section 7.2 presents the system and algorithms used in this chapter, explaining in detail all the steps involved in the disambiguation process. Considerations about the test environment used for evaluation are presented in Section 7.3. The different experiments and the results obtained, as well as a detailed discussion for each of them are described in Sections 7.4, 7.5 and 7.6 respectively. An example of behaviour of

the system and the disambiguation process is presented in Section 7.2.4. Finally, Section 7.7 contains the final conclusions of the chapter.

7.2 System Description

The multilingual technique described in this chapter makes use of the CO-Graph system as source of knowledge for performing WSD. As we briefly introduced above, in this particular version of CO-Graph, we will consider two types of concepts that may eventually become nodes of the co-occurrence graph: First of all, we have specific medical concepts that can be found in the UMLS database, and are identified through their CUIs. As explained in Chapter 6, this identifier is the required output of a system that performs WSD in the biomedical domain, since it unequivocally represents a specific sense. Hence, this information is crucial to exactly determine which sense is the most appropriate for an ambiguous word given its context. However, the UMLS database is mainly restricted to the English language, and hence we need to define another type of concepts which represent the additional information given by other languages, and even by the English language. This second type of concepts are words in the documents, carefully annotated and filtered in order to eliminate all the non-informative words. In particular, we will focus on nouns and adjectives for considering informative words and avoiding introducing too much information into the knowledge base (the co-occurrence graph), which may lead to unmanageable graphs.

Figure 7.1 illustrates the complete system: In part **a)**, we can observe the creation of the knowledge base, which requires a preliminary annotation step. For this step, we have documents written both in English and in any other language which will be used for enriching the knowledge base. The text of each of the documents in the original set is transformed into medical concepts (UMLS CUIs), and nouns and adjectives from English and other languages are extracted. This new document set is then used for building the co-occurrence graph, through the statistical analysis explained in Chapter 3. Part **b)** of the figure represents the disambiguation of a test instance. First, the test instance has to be translated into every language in the multilingual corpus. The ambiguous target term (represented by X in the figure) is located in the text, and its possible senses ($X_1, X_2, ..., X_n$) are extracted from a dictionary. Then, the text of the original test instance, written in English, is mapped onto CUIs. Also nouns and adjectives are extracted from the English sentence, as well as from the sentence translated into all the considered languages. With this information (CUIs and textual information) we can feed the co-occurrence graph and apply a disambiguation algorithm that will select, among those possible solutions, the most suitable sense of the ambiguous term in that context.

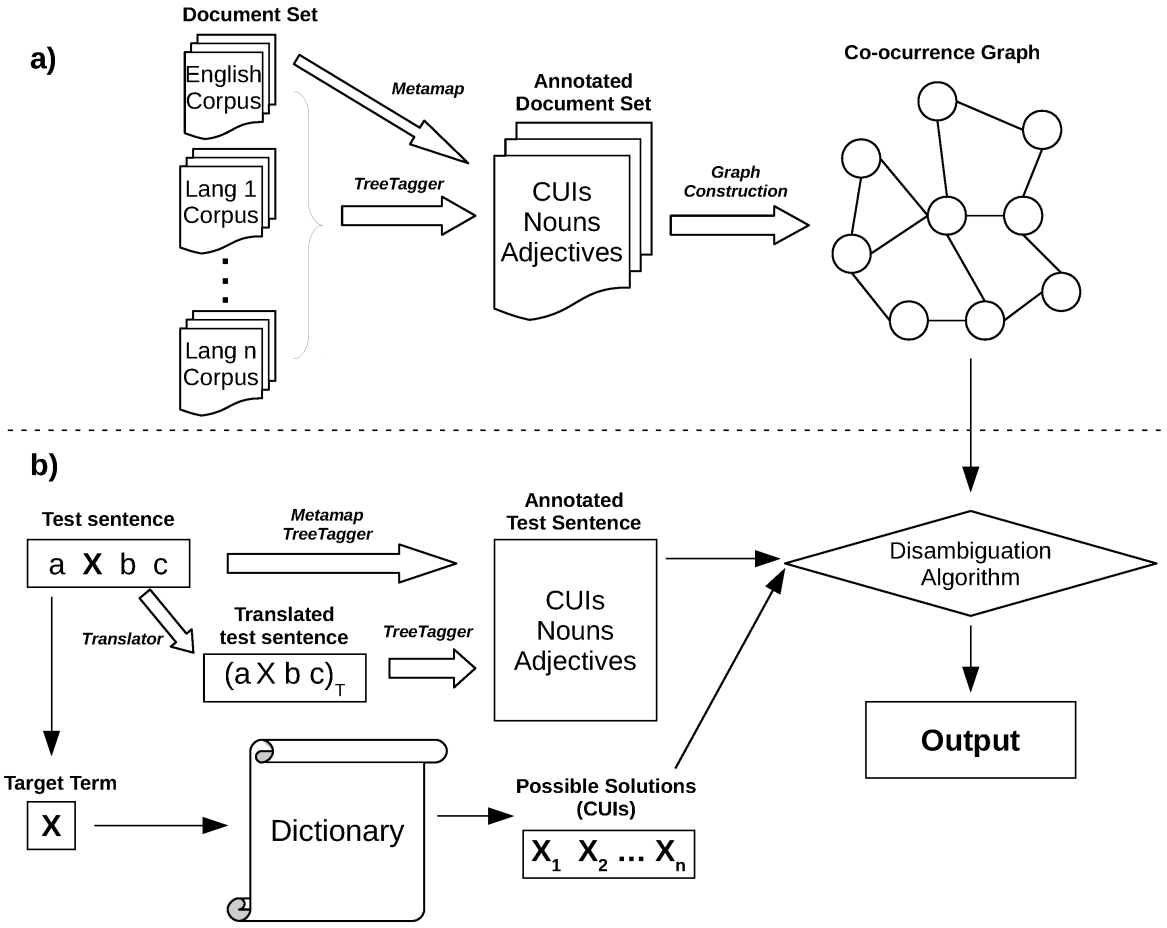


Figure 7.1: Construction of the co-occurrence graph (part **a**) and disambiguation of a test instance (part **b**).

7.2.1 Annotation

As we stated before, the co-occurrence graph will be populated with two different types of concepts: CUIs and words. In order to filter all the non-informative words out of the text documents, we need to lemmatize and tag those documents with Part-Of-Speech (POS) tags. This procedure is automatically performed by the TreeTagger tool (Schmid, 1994), both for English and for the other languages in the multilingual corpus considered, in a similar way to that explained in Chapter 4, in which we tackled the Cross-Lingual WSD task using many different languages. For generating the other kind of concepts that we want to include in our co-occurrence graph (UMLS CUIs), we will use the MetaMap program (Aronson, 2001), using the same configuration explained in Chapter 6, that is, activating

the disambiguation server for the annotation of documents that will be used for building the co-occurrence graph, but selecting only unsupervised algorithms for performing this initial disambiguation, in order to maintain the unsupervised nature of the CO-Graph system. The rest of the configuration parameters of the Metamap program will be set to their default values.

Figure 7.2 shows an example of the annotation step, for the excerpt of an abstract written in English and Spanish, in order to illustrate the differences introduced in this annotation phase, in relation to the annotation phase explained in Chapter 6. We observe the process of annotating the English text with the MetaMap tool for extracting the CUIs. Also, both documents are annotated with TreeTagger for obtaining nouns and adjectives. The final document contains all the concepts that may eventually become nodes of the co-occurrence graph.

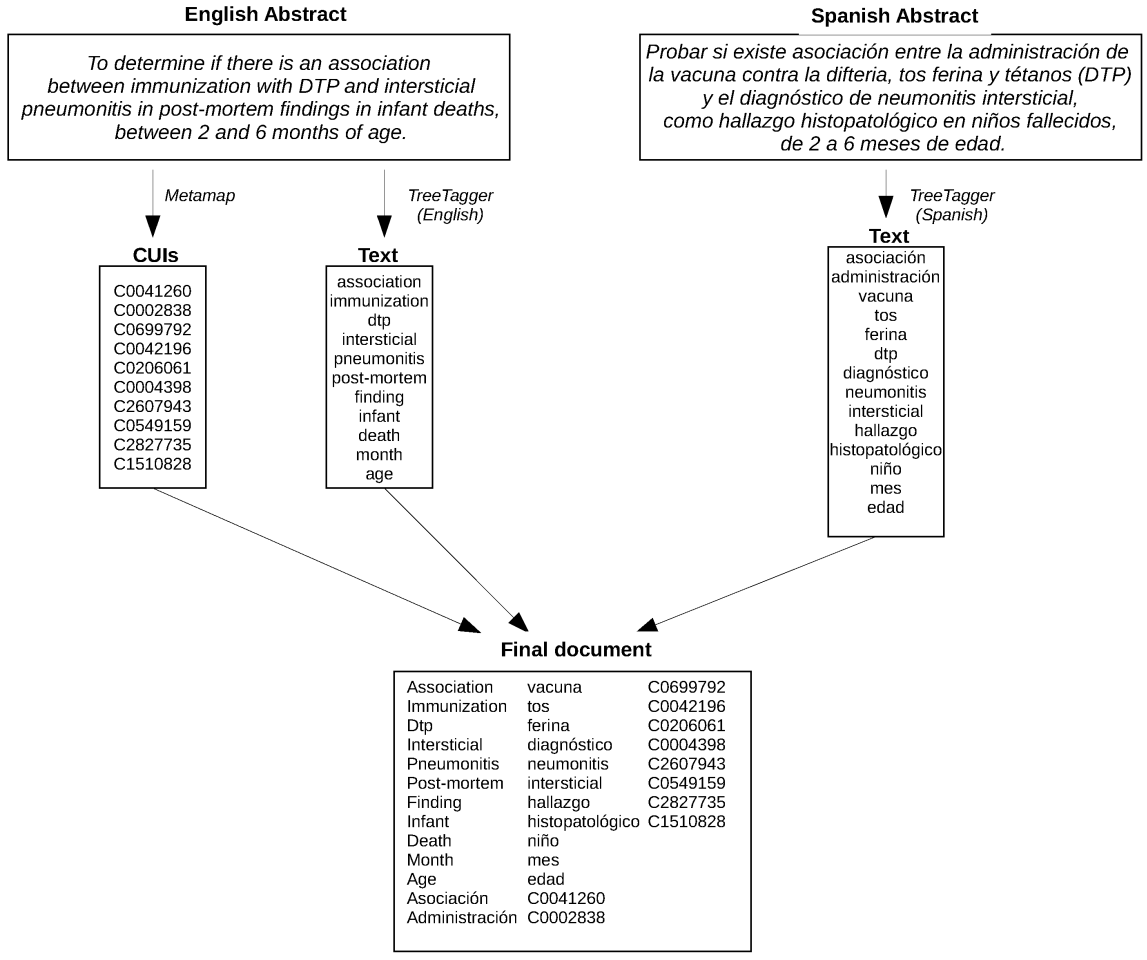


Figure 7.2: Example of annotation of a test instance written in English and Spanish. CUIs from the MetaMap-annotated English document, and nouns and adjectives from both languages are joined together into the final document, which contains concepts for populating the co-occurrence graph.

7.2.2 Graph Construction

The step involving the construction of the co-occurrence graph will follow the process detailed in Chapter 3. In this case, the nodes of the co-occurrence graph can be divided into two different types: CUIs from UMLS, and nouns and adjectives both from English or from the additional language or languages used as source of multilingual information. Despite these differences in the nature of the concepts that will populate the graph, the statistical process followed for building it is the same. Hence, at the end of this process, we will obtain a graph in which we will find co-occurrence links between CUIs, but also between CUIs and words, and even between two words (written either in the same or in different languages).

7.2.3 Disambiguation

As in Chapter 6, the disambiguation algorithm will allow us to select the most suitable CUI for an ambiguous target term. The dictionary that provides the set of possible CUIs for each ambiguous word in a test dataset is also provided within the task itself.

The information needed to feed the graph for performing the disambiguation should be composed by all the possible types of concepts that can be found in the co-occurrence graph, that is, CUIs, nouns and adjectives in English, and nouns and adjectives in any additional language used for enriching the graph. For generating the CUIs related to the biomedical concepts in the context of the target term (English documents), we will also use the MetaMap program. As before (Chapter 6), in this step the disambiguation server is not used, this way assuring that all the possible ambiguity is present in the input with which we feed the co-occurrence graph, that is, no supervision is introduced in the process. The rest of the configuration parameters of MetaMap are set to their default values. For extracting the information related to the plain text (nouns and adjectives in all the languages involved), we first need to obtain a translation of the test instance, which is only written in English. This translation is automatically obtained through the use of the Yandex translator[i], an automatic translation engine which allows users to obtain translations between a large number of languages. Once we have this translation of the test instance, we can run the TreeTagger tool over the English and the translated version and select the nouns and adjectives of both texts, to enrich the original context containing only the CUIs.

In this chapter we are going to explore two different algorithms for disambiguation:

- **One-Step algorithm**: The first disambiguation algorithm makes use of the weights of the links in the co-occurrence graph, as a measure of the importance of a relationship

[i] https://translate.yandex.com

between a particular solution (one of the senses of the target term) and the concepts found in the context of the target term. In a first step we will locate, in the co-occurrence graph, each of the CUIs representing a solution of a test instance, as well as every other concept (CUI or word) in the context of the target term, both in English and in the additional language or languages.

Using the weights of the co-occurrence graph we can rank the possible senses of the target term in the co-occurrence graph given the test instance. Given a test instance T containing a target term t and the terms of its context C, the set of possible senses of the term is represented by $S_t = s_1, s_2, ..., s_n$. For each s_i, we retrieve from the graph the set of concepts $S_C = c_1, c_2, ..., c_m$, which contains the concepts from C that are directly connected to s_i in the graph. We define the weight of a link between a concept $c_k \in S_c$ and a sense $s_i \in S_t$ to be w_{ki}. Hence, the final weight of s_i, denoted by W_i, is computed through the following formula:

$$W_i = \sum_{k=1}^{m} w_{ki}, \tag{7.1}$$

that is, the final weight of s_i is computed by adding up the weights of links between concepts from the context and s_i itself. After computing the weights of every possible sense of the target term, the system will propose the sense with the highest rank to be the most appropriate sense for the test instance.

- **Personalized PageRank**: The second algorithm that we have selected for this step is the Personalized PageRank algorithm, already described in previous chapters of this thesis, especially in Chapters 4 and 6. In this case, the nodes that will be powered up in vector v of the Personalized PageRank formula are those that represent concepts (CUIs, nouns or adjectives in English, and nouns or adjectives in any additional language) that appear in the context of the target term we want to disambiguate. Again, the proposed solution for the test instance will be the node with highest rank according to the algorithm, among the possible CUIs provided by the dictionary (senses of the ambiguous target term).

7.2.4 Example of Disambiguation

In this section a simplified example of how multilingual information can improve the performance of our system is presented. A particular case of disambiguation will be illustrated, by comparing the behaviour of the our co-occurrence graph when we use only English documents for building the graph, and when multilingual (in this case, Spanish) information is added to the graph.

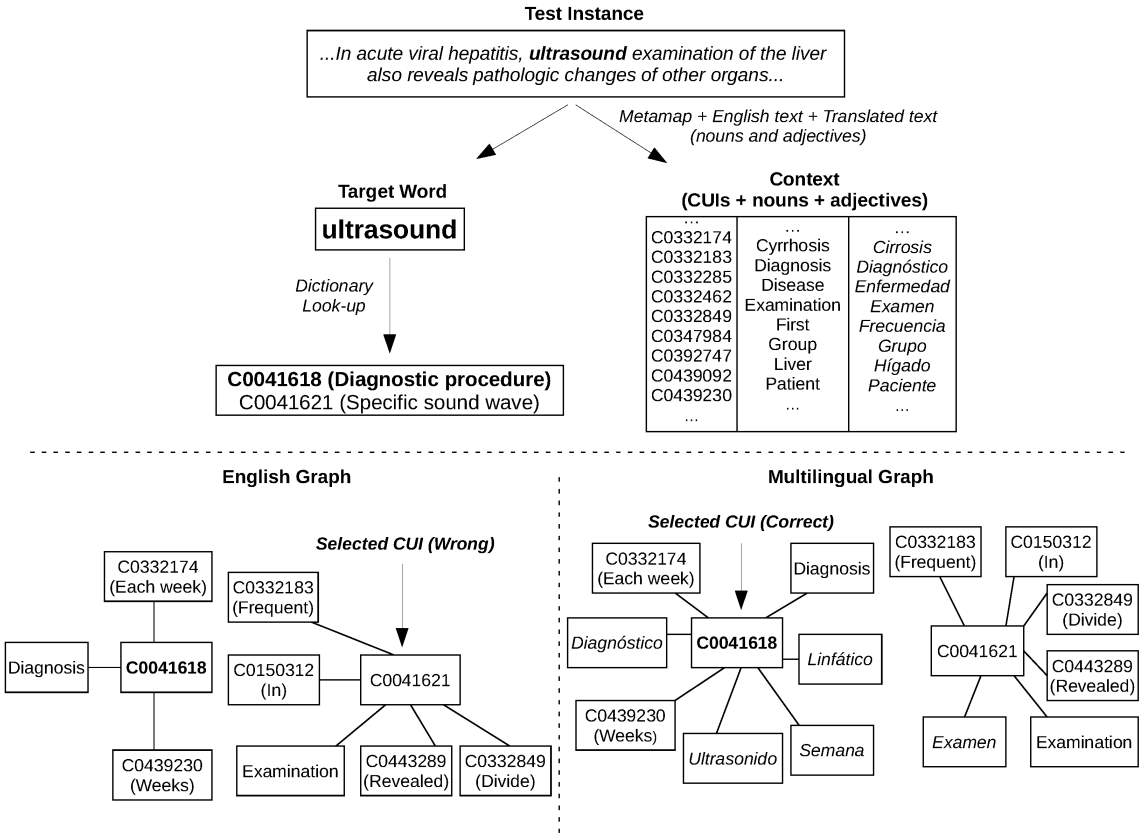


Figure 7.3: Example of disambiguation of a test instance. The top part of the figure shows the annotation of the test instances, while the bottom part compares the behaviour of the English graph and the (English+Spanish) graph.

Figure 7.3 shows this example divided in two parts: the top part of the figure presents a test instance which contains the target word "ultrasound", to be disambiguated. A look-up to the dictionary tells us that the two different senses (CUIs) of "ultrasound" between which our system should discriminate are "C0041618", referred to the process of using ultrasounds for diagnosing a disease, and "C0041621", referred to an ultrasound wave. Through the process of annotation of test instances described in Section 7.2.3, we obtain all the CUIs that represent concepts from the context of the test instance by applying the MetaMap program to the text. Also, nouns and adjectives in both English and Spanish are extracted by running the TreeTagger tool over the original and translated text of the documents. This set of elements represents the input with which we will feed the co-occurrence graph.

The second part of the figure (bottom part) illustrates the differences of applying the disambiguation process using a monolingual English graph, or a multilingual (in this case, English + Spanish) graph. The construction of these two types of graphs will be described in more detail in Sections 7.4 and 7.5. We can observe that the English graph does not classify this instance correctly, while in multilingual graph, the correct sense of "ultrasound"

("C0041618") is selected. If we have a look at the concepts from the context that are directly related to each of the possible senses of the target word, we observe that the English graph contains more concepts related to the wrong sense than to the correct one. That is the reason why in that case, the system selects the wrong CUI ("C0041621"). When we add multilingual information to the graph, the number of related concepts to both the target senses obviously increases. However, in the multilingual graph the sense that now presents more connections with concepts from the context is the correct one ("C0041618"), and hence the algorithm selects that CUI to be the proposed sense for this particular instance.

It is important to notice that in the example we are not explicitly illustrating the use of either of the two disambiguation algorithms studied in this chapter. In both algorithms it is important for a particular sense to have as many direct connections with concepts from the context as possible, in order to be selected as the most appropriate sense for an instance. Nevertheless, there exist other aspects of the algorithms that are also important, such as the weights of the links when it comes to the One-Step algorithm, or the connections between other concepts in the case of the PageRank algorithm.

7.3 Test Environment

In this section we present the framework used in this chapter for testing our system. The test dataset that will be used to evaluate the performance of our system in all the experiments. This test dataset is the NLM-WSD corpus (Weeber et al., 2001), already used as test environment in Chapter 6.

For a proper analysis of the level of improvement that can be achieved by introducing multilinguality in the knowledge base used for disambiguation (the co-occurrence graph), we need to define two types of graphs: The first type is built using English documents, that is, containing CUIs, and nouns and adjectives in English. We will refer to this type of graph as "English graph" in the rest of the chapter. The second type of graph can be seen as an enrichment of the former one, and is built using English documents and documents written in the other language or languages used for adding multilinguality. Hence, this second type of graph will contain CUIs, nouns and adjectives in English, and nouns and adjectives in the other language or languages. This type of graph will be referred to as "Mixed graph" in the rest of the chapter.

As we stated in Section 7.2.2, when we create the co-occurrence graph from a set of documents, we will generate co-occurrence links between two CUIs, but also between a CUI and a word (noun or adjective), or even between words. Therefore, the final structure of the graph (number of nodes and connections between nodes) will change, and we expect this enhanced structure of the graph to improve the accuracy of the system in the WSD task. As we have observed in experiments conducted in previous chapters, the size of the corpus

used for building the graph, and therefore the size of the graph itself, is a very important parameter in this kind of tasks. The methodology that we will follow in all the experiments will be the same: we will compare the overall accuracy achieved by both English and Mixed graphs as we increase the number of documents used for building the graph (knowledge base). This way, we will study whether Mixed graphs built with small subsets of the original multilingual corpus are able to overcome results obtained by English graphs built with larger subsets of the monolingual corpus.

The performance metric used to evaluate system performance in all experiments will be accuracy, as the number of correctly disambiguated instances divided by the total number of instances in the test dataset, expressed in %.

7.4 First Experiment: The EBCRD Corpus

The whole objective of this chapter is to analyse the possible improvements that can be achieved in a Word Sense Disambiguation task when we create a knowledge base with multilingual information. Hence, we will first need a multilingual corpus with documents written in English and at least one more language, in order to create our knowledge base (the co-occurrence graph). As we showed in Figure 7.2, we transform the text documents into documents containing a list of CUIs from the UMLS database, and nouns and adjectives from all the involved languages, that is, we do not take the order of occurrence of the concepts into account. Hence, we do not need to work with parallel corpora in which text is sentence-aligned, but instead we can use comparable corpora containing original documents and their translations into the additional languages.

The multilingual comparable corpus that we have used for this first experiment is the "Elsevier Bilingual Corpus for Rare Diseases" (EBCRD), which we have developed and made publicly available[ii]. It is a bilingual corpus, written in both English and Spanish, and originally created by performing a search for abstracts containing rare diseases (RD) in Ibero-American Elsevier journals whose abstracts are written in both languages and contain at least one term of the NLM-WSD test dataset. This corpus, which contains 94,003 documents per language (for a total of 188,006 documents), will eventually become the knowledge base used for disambiguation.

Once that we have annotated all the documents in the corpus (both those written and English and in Spanish), following the steps described in Section 7.2.1, we are able to build our English and Mixed co-occurrence graphs.

[ii]Corpus available at nlp.uned.es/∼aduque/EBCRD_public.zip

7.4.1 Results

Figure 7.4 shows the results obtained in this first experiment. It illustrates the behaviour of the proposed system when we use English graphs and Mixed graphs for performing the disambiguation, as we increase the number of documents used for building the graph. That is, from the original corpus we take N documents containing each of the possible ambiguous terms.

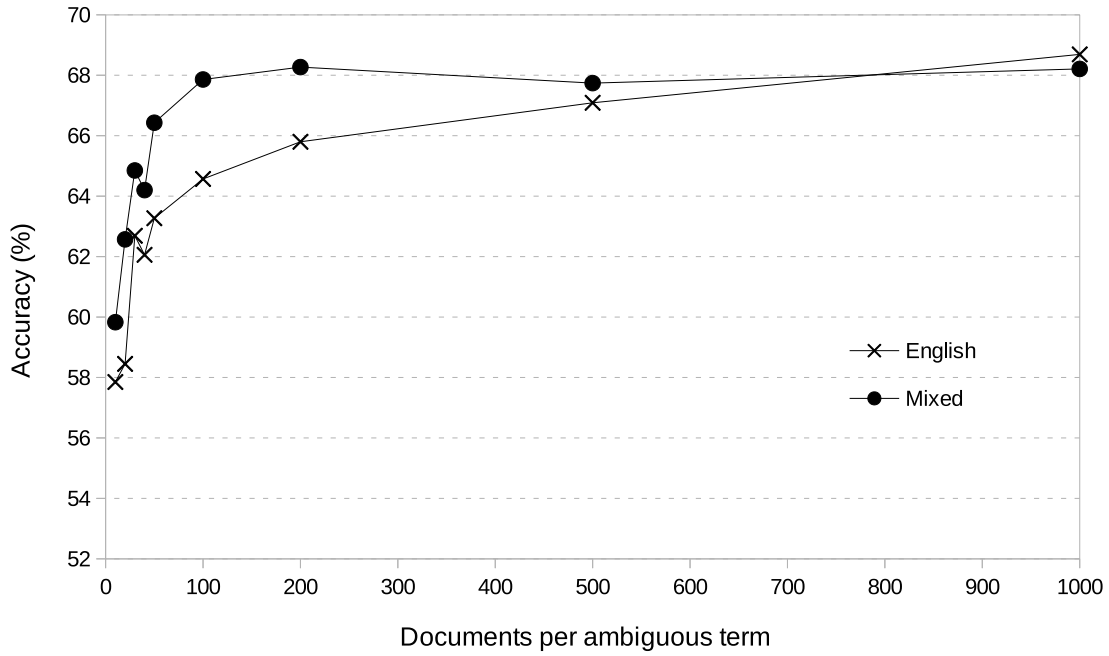


Figure 7.4: Evolution of the accuracy for the English and Mixed graphs of the Elsevier corpus as we increase the number of documents per ambiguous term used for building the co-occurrence graph.

We can observe the improvement achieved when we add the new corpus written in a different language, in this case Spanish, to the co-occurrence graph. Specially, we find the biggest differences when the number of documents used for creating the graph is small, for example for $N = 20$, we get an accuracy of 58.45% for the English graph and 62.57% for the Mixed graph, which represents a relative improvement of 7.05%. When graphs become bigger the improvement achieved by Mixed graphs becomes smaller. Considering what we stated in Chapter 2 and in Section 7.1 about the reduced size of corpora (and specifically multilingual corpora) in the biomedical domain, the fact that multilinguality performs better in smaller datasets is a good indicator.

Results shown in Figure 7.4 refer to the One-Step disambiguation algorithm. We also want to compare the performance of the two considered disambiguation algorithms. Table 7.1 illustrates the accuracy achieved for different sizes of the set of documents used for creating the graphs, both with the One-Step and the Personalized PageRank algorithms.

Docs per term	Total # docs	One-Step	PPR
10	393	59.83	**60.61**
20	779	**62.57**	61.13
30	1,185	**64.85**	62.01
40	1,548	**64.20**	62.47
50	1,908	**66.43**	62.87
100	3,645	**67.86**	67.61
200	6,853	**68.27**	64.75
500	15,202	**67.74**	62.42
1000	26,414	**68.21**	64.47

Table 7.1: Size of the document set (documents per ambiguous term and total number of documents) and comparison (accuracy in %) of the disambiguation algorithms. Bold represents the best disambiguation algorithm in each case.

In this case, we represent both the number of documents per ambiguous word and the total number of documents of the resulting graph. Since documents may contain more than one ambiguous word, the total number of documents used for building the graph is not N times the number of ambiguous words in the dataset (49), but a smaller number, as we can observe in the table. Although there is one case in which the Personalized PageRank (PPR) overcomes the results obtained by the One-Step algorithm, in general we can observe that the latter algorithm generally outperforms PPR.

7.4.2 Discussion

Results shown in Section 7.4.1 give us a first indicator of the benefits of using multilinguality when the number of documents used for building the co-occurrence graph is small. However, as we have stated before, manual translations for creating multilingual corpora are expensive and time consuming. Hence, multilingual corpora are not always available for every subset of documents, specially when the documents are very specific. This leads us to the idea of exploring automatic translations in order to generate multilingual comparable corpora that could be used in a similar way to this experiment. Once we obtain this automatically translated corpus, we will need to analyse whether the quality of the automatic translation allows the system to achieve at least a similar accuracy to the one reported in this first experiment. Table 7.1 has also shown that for this particular corpus, the One-Step algorithm usually performs better than the Personalized PageRank algorithm.

It is important to remark that the multilingual approach offers the best improvements when the co-occurrence graph is built with a small number of documents: between 10 and 200

documents per ambiguous term, that is, between 400 and 7,000 documents in total. When the number of available documents is higher, results converge to similar accuracy values.

7.5 Second Experiment: Automatic Translation

Considering the discussion about the first experiment, the second experiment that we propose is quite straightforward: We want to analyse the performance of the system when we use automatic translations for generating a multilingual corpus, taking an English corpus as original source of information. Many different automatic translators can be found in the literature. In this case we have used the Yandex translator for generating the multilingual documents, since it provides a free API for using the translating services. The Yandex translator is a self-learning statistical machine translation system which creates language models and translation models through the analysis of parallel texts, and connects these models with a decoder. This decoder chooses the best option from the translation model, matches it with the language model to prove its validity, and provides statistics regarding the best result. Using this tool, we generate automatic translations from English to Spanish for every document in the Elsevier Bilingual Corpus for Rare Diseases. This way we are able to compare the performance of our system both using manual and automatic translations from the same original English corpus to enrich the knowledge base (our co-occurrence graph) with multilingual information. The annotation step is followed in the same way as before in order to extract the CUIs and nouns and adjectives in English and Spanish that will populate the co-occurrence graph. We use the same subsets of documents for analysing the evolution of performance as we increase the number of documents per ambiguous term.

7.5.1 Results

Figure 7.5 completes Figure 7.4 with results, for the One-Step algorithm (which performs better than PPR according to Table 7.1), obtained by the system with a Mixed graph created with the original English documents from the corpus, and Spanish documents created with the Yandex translator.

Results obtained by the new Mixed graph (**Mixed Yandex**) also outperform those achieved by the English graph, and even those achieved by the original Mixed graph (**Mixed Manual**). This improvement is particularly noticeable for small subsets of documents. For example, if we consider $N = 30$ (being N the number of documents per ambiguous term), we can observe an accuracy of 62.69% for the English graph, 64.85% for the Mixed Manual graph and 65.98% for the Mixed Yandex graph. That is, the Mixed Manual graph obtains a relative improvement of 3.45% over the English graph, and the Mixed Yandex graph a relative

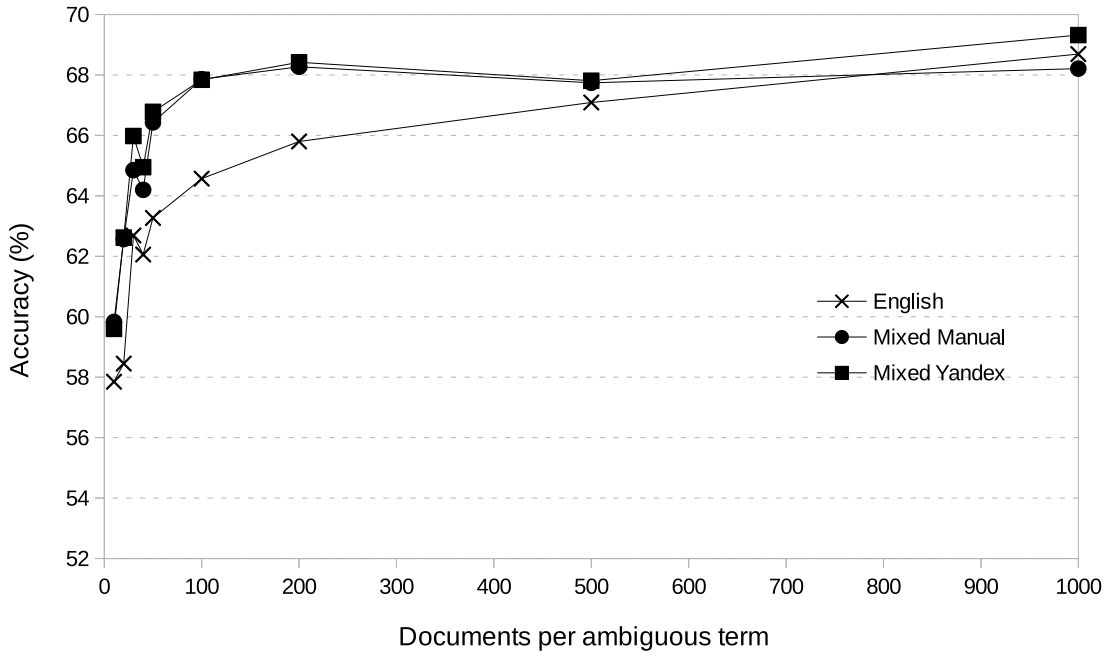


Figure 7.5: Evolution of the accuracy for the English and Mixed graphs of the Elsevier corpus (**English** and **Mixed Manual**) and the Mixed graph built with its Yandex translation (**Mixed Yandex**) as we increase the number of documents per ambiguous term in the knowledge base.

improvement of 1.74% over the Mixed Manual graph (and 5.25% over the English graph). The Mixed Yandex graph is even able to overcome the English graph for bigger subsets of documents, for example for $N = 1000$, the Mixed Yandex graph obtains an accuracy of 69.32% and the English graph an accuracy of 68.69% (relative improvement of 0.92%).

An experiment has been performed for testing the statistical significance of the differences in the values reported in Figure 7.5 (differences between "Mixed Manual" and "English", between "Mixed Yandex" and "English", and between "Mixed Yandex" and "Mixed Manual"). The method followed for comparing the differences between the experiments has been to calculate the achieved accuracy as we increase the number of documents used for building the co-occurrence graph, and then running the statistical significance test against the reported collection of accuracies for each method. As the population cannot be assumed to be normally distributed, in this case we have used a Wilcoxon Signed-Rank test (Wilcoxon, 1945), with a significance trust value of 95%.

Table 7.2 contains those values of accuracy as we increase the number of documents per term, for the "English", "Mixed Manual" and "Mixed Yandex" experiments, using the One-Step algorithm:

With those values, the Wilcoxon Signed-Rank test offers the following values:

Docs per term	English	Mixed Manual	Mixed Yandex
5	56.74	59.83	58.80
10	57.85	59.83	59.60
15	56.94	59.40	60.38
20	58.45	62.57	62.62
25	59.63	63.44	64.05
30	62.69	64.85	65.98
35	62.49	64.75	65.88
40	62.06	64.20	64.95
45	63.02	65.83	66.33
50	63.27	66.43	66.78

Table 7.2: Accuracy (in %) achieved by the English, Mixed Manual and Mixed Yandex graphs as we increase the number of documents per ambiguous term (column **Docs per term**) used for building the co-occurrence graph.

- The difference between columns "Mixed Manual" and "English" is **statistically significant**, since $p - value = 0.00512$.
- The difference between columns "Mixed Yandex" and "English" is **statistically significant**, since $p - value = 0.00512$.
- The difference between columns "Mixed Yandex" and "Mixed Manual" is **not statistically significant**, since $p - value = 0.07508$.

As we can observe, we have included in the table those values for which we are interested in knowing the statistical significance, that is, the accuracy achieved by graphs built with a small number of documents. In that range, we can conclude that using multilingual information (obtained from either manual or automatic translations) allows us to obtain significantly better accuracy in the test dataset. The difference between working with manual and automatic translations is not so relevant for the task.

Considering that automatic translations are far easier to obtain than manual translations, we have performed an additional experiment in which we obtain translations for a small subset of documents in other languages apart from Spanish. Table 7.3 shows the results obtained by the system for the English graph, and for Mixed graphs created with different combinations of languages. We have selected a subset of 50 documents per ambiguous word to analyse the results, since previous experiments have shown that multilinguality is able to get better improvements for smaller subsets of documents. The considered languages are: Spanish (SP), German (GE), Russian (RU) and Italian (IT). We have selected those particular languages in order to analyse the influence of languages with different roots. In this case, Spanish and Italian are Romance languages, whereas German is a Germanic language with a completely different origin and structure. We have also selected the Russian language (an

East Slavic branch of Indo-European languages) since Yandex is a Russian company and the Russian language was one of the first languages in the Yandex translator.

Language(s)	Accuracy (%)
EN	**63.27**
EN+SP	**66.78**
EN+GE	64.75
EN+RU	65.45
EN+IT	65.50
EN+SP+GE	**67.76**
EN+SP+IT	67.26
EN+SP+RU	67.49
EN+GE+RU	65.91
EN+GE+IT	66.11
EN+RU+IT	66.01
EN+SP+GE+RU	**67.71**
EN+SP+GE+IT	67.66
EN+SP+RU+IT	66.98
EN+GE+RU+IT	66.51
EN+SP+GE+RU+IT	**67.24**

Table 7.3: Results (accuracy in %, 50 documents per ambiguous word) obtained with combination of different languages: English (EN), Spanish (SP), German (GE), Russian (RU) and Italian (IT). Bold highlights the best results for the combination of English with none, 1, 2, 3 or 4 additional languages.

We can observe results for the English graph, and Mixed graphs created with the combination of English documents and one, two, three or all of the considered additional languages. Spanish is the language that obtains better results when combined alone with English, followed by Italian, while Russian and German offer less improvement when combined with English. This fact may indicate that the translator is working better for Romance languages. However, the best result (accuracy of 67.76%, relative improvement of 7.10% over the English graph) is achieved when we combine Spanish and German translations with English documents. This can be due to the differences between the root languages of Spanish and German (Latin and Germanic languages respectively). The amount of information introduced by two languages of different origins and structured is probably higher than what we can expect from two similar languages, such as Spanish and Italian together.

Statistical tests have also been applied in a similar way as described in Table 7.2, although we have only considered those cases interesting for our purposes. In particular, we have tested the significance of the differences between results obtained only using monolingual information (row **EN** in the Table), adding multilingual information in Spanish (row **EN+SP**), since those are the conditions for previous and subsequent experiments, and adding multilingual information in Spanish and German (row **EN+SP+GE**), as that is the case in which the highest accuracy is achieved.

Table 7.4 of the present document contains the values of accuracy for the three rows above-

mentioned:

Docs per term	EN	EN+SP	EN+SP+GE
5	56.74	59.83	57.04
10	57.85	59.83	59.00
15	56.94	59.40	59.33
20	58.45	62.57	60.18
25	59.63	63.44	60.51
30	62.69	64.85	65.50
35	62.49	64.75	67.31
40	62.06	64.20	69.19
45	63.02	65.83	68.09
50	63.27	66.43	67.76

Table 7.4: Accuracy (in %) achieved by graphs built with only English information (Column **EN**), English and Spanish information (Column **EN+SP**) and English, Spanish and German information (Column **EN+SP+GE**), as we increase the number of documents per ambiguous term (column **Docs per term**) used for building the co-occurrence graph.

The two first columns are equal to those in Table 7.2, so the statistical significance between them has already been tested. The Wilcoxon Signed-Rank test results are:

- The difference between columns "EN+SP" and "EN" is **statistically significant**, since $p - value = 0.00338$.
- The difference between columns "EN+SP+GE" and "EN" is **statistically significant**, since $p - value = 0.00512$.
- The difference between columns "EN+SP+GE" and "EN+SP" is **not statistically significant**, since $p - value = 0.96012$.

As we had already proven, the differences between using multilingual information and not using it are statistically significant. In this case, we also prove this fact regarding Spanish and German as languages providing the multilingual enhancement. However, we can also observe that there is not statistical significance between the two multilingual scenarios considered, that is, one enriched only with Spanish words, and other enriched with both Spanish and German words. As in Table 7.2, we have computed these significances within a range of small subsets of documents used for building the co-occurrence graphs (up to 50).

7.5.2 Discussion

Although current machine translation systems are not able of outperforming manual translations, in our case we observe that the results obtained with a multilingual knowledge

base automatically created are better than those results obtained with manually translated documents. The main reason why this could be happening is the nature of the disambiguation system described and the WSD task we are facing. Apart from CUIs and nouns and adjectives from the original English documents, we are only using nouns and adjectives from the translated texts for building our knowledge base. We consider that it is likely that our system is giving more importance to the correct translation of these words than to the structure of the translated sentences (which is far more difficult for an automatic translator to represent correctly). Moreover, sometimes manual translations rely on the personal interpretation of the human translator, which can lead to less literal translations than those obtained by an automatic system. Although this can be positive when we expect a more thorough translation, in this case the creation of more literal translations can be beneficial for our purposes, since they are more likely to directly solve the ambiguity of some words. In this second experiment, we ascertain that using a small number of documents (between 10 and 200) still offers the best improvements when it comes to a multilingual approach, in this case obtained through automatic translations.

Besides, Table 7.3 also indicates that generating automatic Spanish translations with Yandex can offer successful results, while the combination of other languages may only slightly improve the results when the languages are different enough to provide new information.

7.6 Third Experiment: NLM Corpus

In the previous experiments we have proven the usefulness of applying multilinguality for performing WSD in the biomedical domain. However, as we stated in Section 7.4, the multilingual corpus used in those experiments was created from a search related to rare diseases. The information obtained in the second experiment regarding the possibility of using automatic translations allows us to explore more specific corpora, a priori written only in English, which could offer better results. In particular, as the NLM test dataset is generated from PubMed[iii] abstracts, we are interested in using a corpus with abstracts from PubMed which contain ambiguous terms from the NLM test dataset. This way, we expect our knowledge base (the co-occurrence graph) to get closer to the characteristics of the test dataset, and hence achieve better accuracy. The last step will be to analyse whether a graph enriched with automatic translations of this "NLM-related" corpus is able to improve the accuracy of the system in the proposed WSD task, in a similar way to the previous experiments.

The NLM-related corpus that will be used in this experiment was described in Section 6.3.2 of Chapter 6, and is composed of 35,282 abstracts from PubMed, each of them containing at

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

least one ambiguous term from the NLM-WSD test dataset[iv].

We applied the same procedure explained in Section 7.5 to this NLM-related corpus, using Yandex translator for generating the Spanish translation of each document in the English corpus. Then we perform the annotation step for extracting CUIs and nouns and adjectives in both languages, and we create the final co-occurrence English and Mixed graphs. As we did in the previous experiments, we are going to analyse the performance of the system as we increase the total number of documents used for building the graphs. In this case, the number of documents per ambiguous term is already balanced, and hence a simple random subsampling of the full corpus should be enough to obtain subsets of documents in which we find a similar number of documents per ambiguous term.

7.6.1 Results

Table 7.5 shows the results obtained by the English and Mixed graphs built with documents from the NLM-related corpus. Performance by both One-Step and PPR algorithm is also shown, to analyse whether they behave differently when graphs are built from this new corpus.

	One-Step		PPR	
Total # docs	**English**	**Mixed**	**English**	**Mixed**
1K	73.14	**74.62**	66.28	66.98
10K	74.42	**74.67**	73.61	74.52
20K	75.55	76.53	74.77	**76.90**
Full	76.05	77.48	**77.68**	77.63

Table 7.5: Results (accuracy in %) using the NLM-related corpus, for different sizes of the document set used for building the graph. Bold highlights the best configuration (algorithm and type of graph) for each experiment.

Overall accuracy obtained in this experiment is quite higher even for English graphs, probably due to the similarities between the test dataset and the NLM-related corpus used for building the graphs. Both the test dataset and the co-occurrence graph are created with abstracts downloaded from PubMed, and hence knowledge in the co-occurrence graph is more likely to present the same characteristics as the test dataset, which may lead to better results in the disambiguation process. Despite this improvement of the general results, we can observe in the table that Mixed graphs are still able to overcome English graphs, although the differences are smaller. These differences are also more important when we consider small subsets of documents (relative improvement of 2.02% for graphs built from 1,000 documents), while English and Mixed graphs perform similarly when we use the complete set of 35,282 documents for building them. Differences between the disambiguation algorithms

[iv] Corpus available at nlp.uned.es/∼aduque/NLM_related_public.zip

are also bigger as the subset of documents is smaller, specially for 1,000 documents (relative improvement of 11.41% for Mixed graphs). However, as the number of documents increases both algorithms also present similar results.

7.6.2 Comparative

Finally, we want to compare in Table 7.6 the best performance achieved by our multilingual system with results offered by other state-of-the-art unsupervised systems performing WSD in the biomedical domain. For this comparison, we take from Table 7.5 the best accuracy obtained by a Mixed graph which still present differences with the English graph of the same subset of documents. In this case, the only additional language used for the Mixed graph is Spanish, that is, concepts in the graph are CUIs, nouns and adjectives in English, and nouns and adjectives in Spanish. This best result of 76.90% of accuracy is achieved by a Mixed graph created with 20K documents from the NLM-related corpus, selecting PPR as disambiguation algorithm, and 76.53% of accuracy under the same conditions, selecting One-Step as disambiguation algorithm.

System	NLM Test Dataset
MetaMap Baseline	49.13
PPR+UMLS	68.10
AEC	68.36
JDI	74.75*
MRD	63.89
2MRD	55.00
CO-GRAPH (One-Step)	76.53
CO-GRAPH (PPR)	**76.90**

Table 7.6: Comparison of the accuracy (%) achieved by state-of-the-art unsupervised systems (see text), and our multilingual co-occurrence graph-based system (CO-Graph). The first row corresponds to a baseline showing the performance of the MetaMap disambiguation server over the test dataset. The asterisk in row **JDI** indicates modifications in the test dataset (see text).

In the first row, we show results obtained by running the MetaMap program against the test dataset, and making use of the disambiguation server under the same conditions we used for annotating the documents when building the co-occurrence graph, as explained in Section 7.2.1. As we can observe, these results are quite low in comparison with the accuracy achieved by our system in all the experiments reported in this Chapter. Results from our system are then compared against different WSD systems. All these system, whose characteristics have been already described in Chapter 2 and in Section 6.4.2 of Chapter 6, are monolingual, that is, they do not make use of multilingual information for enriching the available knowledge. As we can observe in the table, our system outperforms all the state-of-the-art knowledge-based and unsupervised methods when applied to the NLM dataset, and even semi-supervised ones.

7.7 Conclusions

In this chapter we have presented an adaptation of our unsupervised CO-Graph system for performing Word Sense Disambiguation in the biomedical domain. The objective of the study is to determine whether multilingual information is able to improve the results obtained by monolingual approaches in WSD tasks, and under which conditions this improvement is real and significative. In three different experiments performed over a test dataset widely used in the literature, multilinguality has been proven useful for WSD, particularly when the knowledge base is limited, that is, the number of documents used for building the graph is small. We have used the NLM-WSD test dataset, composed of general ambiguous words in biomedicine, in order to be able to analyse the performance of our system when varying the amount of available information for building the co-occurrence graph. Results obtained using two different corpora for building the graphs, one unrelated and the other related to the test dataset, indicate that a big corpus unrelated to the test dataset achieves worse results than a small corpus, but related to the test dataset (for example, the NLM-related corpus with only 1,000 documents, that is, around 20 documents per ambiguous term). These facts lead us to extrapolate the results obtained in the experiments, and consider that multilinguality would also be useful when considering ambiguous words for which less occurrences could be found in the literature (for example, terms for which one of their senses represented a rare disease poorly documented).

We have observed how smaller sizes of the co-occurrence graph lead to similar or even better results than those obtained with bigger graphs, which is a very good indicator in terms of efficiency and resource consumption. For example, we can observe in Table 7.5 that Mixed graphs built from a subset of 1,000 documents, whose size is approximately 40K nodes and 2 million links, are able to obtain similar results to English graphs from a subset of 20,000 documents, containing 200K nodes and more than 8 million links. The obtained improvements suggest that the translation of general terms of the context of an ambiguous term provides an important source of information to select the correct biomedical concept associated to that ambiguous biomedical term. This information can be eventually transformed into structured knowledge that allows us to disambiguate the biomedical terms in the test instances.

Automatic translations, which are normally much easier to obtain than manual translations, are able to match, and even outperform results from manual translations. This makes the approach proposed in this chapter highly suitable for this kind of tasks, due to the lack of multilingual corpora in many scenarios of the biomedical domain. When using automatic translations, additional languages are proven to be more useful for WSD when their differences with the original language (in this case, English) are bigger. In general, the addition of new languages to the multilingual co-occurrence graph only improves the overall results when those languages are different enough to provide new information.

8

CONCLUSIONS AND FUTURE WORK

In literature and in life we ultimately pursue, not conclusions, but beginnings.

Sam Tanenhaus

Contents

This chapter represents the conclusion of the thesis. We will summarize the main contributions generated along the development of the thesis, remarking those techniques, datasets and solved issues that are particularly innovative within the field of study (Section 8.1). The main conclusions of the work presented here will be detailed in Section 8.2, in the form of answers to the Research Questions that we proposed in Chapter 1. Through these answers and conclusions we intend to determine whether the main Research Objective has been fulfilled. In Section 8.3 we will draw some lines of research derived from this thesis which we want to explore in future studies. Finally, we present the publications derived from the development of this thesis in Section 8.4.

8.1 Main Contributions

In this first section, we will summarize the main contributions developed through the research process described in this thesis:

- **CO-Graph technique**: In Chapter 3 we described the mathematical foundations of the technique that has been used along all the research process. This technique has been compared to other statistical models found in the literature for extracting co-occurrence information. Then, results from Chapters 3 and 4 indicate that our method is able to achieve better performance in Word Sense Induction and Word Sense Disambiguation tasks.
- **Dictionary study**: We have performed an exhaustive study on the comparison of bilingual dictionaries and the implications derived from their use within the pipeline of a system performing CLWSD. As far as we know, no similar study can be found in the literature for its application to CLWSD tasks.
- **Biomedical WSD**: We have adapted our technique based on co-occurrence graphs for its application to the biomedical domain, and more particularly on biomedical WSD tasks. For this aim, we have studied the UMLS biomedical database and we have adjusted the process of building the graphs for using identifiers as nodes, instead of raw text. Results obtained using two widely studied test datasets indicate that our technique can be successfully extrapolated for performing WSD in this specific domain.
- **Mono and multilingual corpora**: We have addressed the creation of comparable corpora in Chapter 7, in order to overcome the lack of this kind of corpora in the biomedical domain. More particularly, we have created the Elsevier Bilingual Corpus for Rare Diseases (EBCRD), and English-Spanish comparable corpora containing medical abstracts written in both languages. Also, we created the NLM-related corpus, which contains abstracts originally written in English. Automatic translations

from subsets of the NLM-related corpus have also been generated in different target languages for testing their usefulness for including multilingual information to our knowledge base.

- **Multilinguality in biomedical WSD tasks**: Apart from the creation of multilingual comparable corpora, which have been very helpful in the development of the work described in Chapter 7, the study on multilinguality and its potentiality for improving biomedical WSD tasks is a very important contribution itself. As far as we know, there are no similar studies in the literature, and hence the analysis developed in that chapter can be taken as a starting point for future lines of work regarding multilingual WSD in the biomedical domain.

8.2 Answers to Research Questions

In Chapter 1, we stated the Research Objective of this thesis:

RESEARCH OBJECTIVE: ***Study the problem of Word Sense Disambiguation in the scope of Natural Language Processing, and the importance of solving this problem also in specific domains such as biomedicine. Analyse the usefulness of multilinguality to improve systems performing WSD and develop an unsupervised graph-based system able to overcome state-of-the-art techniques in different WSD tasks.***

Then, this main goal was divided into nine different Research Questions, each of them representing a smaller objective that would lead us in the decisions taken along the development of the thesis. In this section, we want to offer our answers to those Research Questions, as a way to summarize the main conclusions of this work, regarding the different studies carried on and described in previous chapters.

Research Question 1: *Considering the idea of coherence inside a document, is the co-occurrence graph a valid structured representation of the information inside a corpus, for addressing Word Sense Disambiguation tasks?*

We have proved that our proposed technique CO-Graph is able to achieve similar and better results compared to other unsupervised and knowledge-based techniques when tackling WSD tasks such as CLWSD and biomedical WSD. In Chapter 3 we showed how the statistical model used for determining the significance of each co-occurrence is better than other models based on the Chi-Squared test or on the G-Test. This fact leads us to conclude that the hypothesis of coherence inside a document is valid for representing the information inside a corpus in a structured way and subsequently use this structured information for performing WSD tasks.

Research Question 2: *Once that the information has been formally represented in a co-occurrence graph, which disambiguation algorithm makes better use of this structured source of knowledge to perform disambiguation?*

We have tested many algorithms in the disambiguation phase of the proposed system: in the CLWSD task we applied community-based algorithms such as Walktrap and Chinese Whispers, as well as the Dijkstra's algorithm and two versions of PageRank (basic and with priors). In that work, we showed that the differences between the algorithms used for disambiguation were not high, although algorithms making use of the weights of the links in the graph (community-based and Dijkstra's) appeared to offer slightly better results than those only based on the structural information of the graph (PageRank and Personalized PageRank). This conclusion is also obtained after the study performed in Chapter 7, in which the One-Step algorithm, which makes use of the weights of links in the graph, performs better than the Personalized PageRank algorithm in most of the cases. Hence, the main conclusion to this Research Question would be that once we have the structured representation of the information in the graph, it is advisable to use disambiguation algorithms that take into account all the information offered by the graph, that is, both structural information (links between concepts) and quantitative information (weights of those links). The algorithm offering the best results, amongst those using all the information of the graph, may depend on the particular WSD task.

Also, in Chapters 4 and 5 we observed that the use of prior information regarding the translation probabilities of ambiguous target words is very useful for performing the final disambiguation, so it is also highly recommended for a system performing WSD to include, when possible, this kind of information within its pipeline. For example, in those chapters we were able to use those translation probabilities without compromising the unsupervised nature of our system, through the use of the GIZA++ statistical aligner.

Research Question 3: *How close can get the results obtained by an unsupervised Word Sense Disambiguation algorithm to those achieved by a supervised one?*

As we introduced in Chapter 1, two different WSD main tasks have been tackled along the development of this thesis in order to prove the validity of the CO-Graph system. Comparisons with state-of-the-art systems have been presented for both tasks, more particularly in Tables 4.7, 5.8 and 5.9 regarding CLWSD, and Tables 6.3 and 7.6 regarding mono and multilingual WSD in the biomedical domain. Most of the systems to which we have compared our results are unsupervised and knowledge-based, in order to maintain a fair comparison given the characteristics of our own technique. However, especially in the field of CLWSD we have also shown results from supervised systems. Although most of the studied supervised systems that have tackled the same tasks that we have are able to outperform our results, there are some cases in which the CO-Graph system is able to obtain similar results and even overcome supervised systems. For example, in the SemEval 2010 competition the CO-Graph system is able to get better results than the supervised system presented by van Gompel

(2010) in the Spanish and Dutch Out-Of-Five evaluation schemes, and those achieved by Vilariño et al. (2010) in the Spanish Out-Of-Five evaluation scheme. In the SemEval 2013 competition, the best supervised techniques (WSD2 and HLTDI, proposed by van Gompel and van den Bosch (2013) and Rudnick et al. (2013), respectively) obtain better results than CO-Graph. However, the differences between those systems are not really high (around 3% to 5% better for most of the studied languages).

When it comes to the biomedical domain, supervised works can be found in the literature obtaining quite higher results compared to those obtained by knowledge-based techniques. For example, in the work by Stevenson et al. (2009), results for the Acronym test dataset offer an accuracy of more than 97%, while supervised systems are able to reach an accuracy of 91% for the NLM-WSD test dataset (McInnes and Stevenson, 2014).

The differences between supervised and unsupervised (or knowledge-based) approaches are then much higher in the case of WSD in the biomedical domain than those in CLWSD tasks. The main reason for this, is because the supervised methods addressing the CLWSD task cannot be considered as fully supervised, since the amount of instances annotated with the correct translations is quite reduced. Hence, the supervised systems rely on the creation of automatic annotations or labels for the training instances, so they can be considered to be semi-supervised systems, following the same ideas behind the classification described in Section 2.3 of Chapter 2. However, the supervised systems found in the biomedical domain are fully supervised, and hence they are able to obtain much higher accuracies, far from those obtained by knowledge-based systems like ours.

Research Question 4: *How can we combine multilingual information available in parallel and comparable corpora with our proposed technique for performing Cross-Lingual Word Sense Disambiguation?*

As we have observed in Chapters 4 and 5, the Cross-Lingual WSD task asks the system to offer the most suitable translations in a target language, for an ambiguous word initially written in a source language. Regarding this required output, we need to include, at least, both information of the source and target languages in the pipeline performing the final disambiguation. Although some systems generate data structures which include joint information about both languages, and even about all the considered languages in the task at the same time (Silberer and Ponzetto, 2010), most of the systems performing CLWSD use a pipeline which separates the disambiguation and the use of multilingual information in different steps. This way, one step of the overall process is more focused on translation issues, and the other step takes care of the disambiguation itself. The language in which the final disambiguation is performed can be the source language (Apidianaki, 2013), or the target language (Guo and Diab, 2010; Tan and Bond, 2013), as we explained in Chapter 2. We have followed the latter scheme for implementing our disambiguation pipeline, performing translation operations through the statistical aligner GIZA++ in one step, and relying on a graph built with information from the target language for disambiguation issues. Results obtained in both SemEval 2010

and 2013 competitions show that this strategy for combining multilingual information from parallel corpora with our technique based on co-occurrence graphs is able to overcome state-of-the-art unsupervised systems and even supervised systems in some cases.

Research Question 5: *Is our unsupervised graph-based technique able to overcome other state-of-the-art approaches in Cross-Lingual Word Sense Disambiguation tasks?*

Results in Chapter 4 clearly show that our unsupervised graph-based technique CO-Graph is able to outperform those results offered by unsupervised systems participating in the SemEval 2010 and 2013 CLWSD tasks in most cases. In order to offer an accurate answer to this Research Question, it is important to consider the five different target languages involved in the CLWSD tasks (Spanish, French, Italian, German and Dutch), as well as the two evaluation schemes (Best and Out-Of-Five). In particular, compared to the best unsupervised system in each case, our system obtains better results for German, Italian and Dutch in the Best scheme of 2010, for Spanish, Italian and Dutch in the Out-Of-Five scheme of 2010, for German, French and Dutch in the Best scheme of 2013 and for the five languages in the Out-Of-Five scheme of 2013. In the remaining cases, our system usually obtains the second best results. However, it is also important to remark that all the systems that are able to outperform ours in some cases make use of external resources apart from the Europarl corpus provided by the organizers of the task for performing the disambiguation, which is the only resource used by CO-Graph. In 2010, this system is called T3-COLEUR (Guo and Diab, 2010), which makes use of WordNet synsets for performing WSD in English and then use the selected sense for finding the most appropriate translation, according to the probabilities given by GIZA++. In 2013, these systems which perform better than ours in some cases are LIMSI (Apidianaki, 2013), using the JRC-Acquis corpus as an additional resource, and NRC-SMT (Carpuat, 2013), which adds information from news data extracted from previous SemEval competitions to the source of information.

Research Question 6: *What is the impact of selecting different bilingual dictionaries in a system performing Cross-Lingual Word Sense Disambiguation?*

The impact of choosing or generating different bilingual dictionaries for CLWSD task has been studied in Chapter 5. In that study, four different dictionaries were analysed: a generic bilingual dictionary manually created, a collaboratively created dictionary, extracted from the Multilingual Central Repository (Atserias et al., 2004), a dictionary extracted from the BabelNet multilingual network (Navigli and Ponzetto, 2010), which can be seen as semi-automatically created, and finally a statistical dictionary created in an automatic unsupervised way through the use of GIZA++. This last dictionary offers the best results in terms of coverage (number of translations from the source language to the target one which are also contained in the Gold-Standard) and precision, inside a disambiguation pipeline that makes use of our CO-Graph system. The use of prior translation probabilities offered by GIZA++ is a key step for drastically improving the performance of a system using this

kind of statistical dictionaries, over the other analysed dictionaries. The main drawback that can be found when using GIZA++ is the high number of translations that it offers for each word, even when the intersection between the two languages is extracted (only those translations presenting a reverse equivalent translation are considered). This fact usually leads to a worsening in the performance offered by the system, not only because of the high number of possible translations for each target word, but also because of the noise introduced when translating words from the context to all their possible translations offered by GIZA++. For solving this issue, the bilingual dictionary generated with GIZA++ needs to be pruned, only maintaining those translations with a high probability of occurrence. In our case, we found that pruning the dictionary up to ten translations per word offers successful results.

Research Question 7: *Can we successfully apply our original disambiguation technique to Word Sense Disambiguation in the biomedical domain? Which are the adjustments that should be applied to our algorithm for tackling the WSD task in this domain?*

For answering this research question, we focused on the biomedical WSD task described in Chapters 6 and 7, in which the correct sense of a biomedical ambiguous term or acronym has to be found given its context, among all the possible senses to which this term can refer. The main adjustment that has to be done to our co-occurrence graph technique is related to the nature of those concepts that may eventually become nodes of the co-occurrence graph. The senses to which an ambiguous concept may refer are represented as identifiers called CUIs (Concept Unique Identifiers), which are defined in the UMLS database, possibly the most important structured source of knowledge when it comes to biomedical information in English. Hence, in a similar way to what we did in the CLWSD task, in which we populated the graph with the final concepts that could become the output of a given test instance (in that case, words in the target language), this time we need to populate the graph with those CUIs referring to biomedical concepts. For that purpose, an additional annotation step had to be done, for transforming raw text extracted from abstracts from biomedical journals into unique identifiers that helped us to perform the disambiguation.

Considering that the biomedical WSD task is monolingual, that is, there is no need for performing translations or using multilingual corpora for finding the most suitable sense for an ambiguous concept, we first tackled the task from that monolingual perspective, in order to analyse whether our technique was able to offer promising initial results. In that study, described in Chapter 6, we show how our technique is able to offer results that overcome state-of-the-art systems in the task, particularly for the NLM-WSD test dataset (Weeber et al., 2001), and also competitive results in the Acronym test dataset (Stevenson et al., 2009). This way, we obtain an answer for this Research Question by proving that the developed CO-Graph technique, properly adjusted in order to fulfill the requirements of this slightly different task, can be successfully applied to solve WSD problems in the biomedical domain.

Research Question 8: *Can unsupervised Word Sense Disambiguation be improved by multilingual information in the biomedical domain? Under which circumstances?*

Chapter 7 is mainly devoted to answering this particular Research Question. The main objective of that chapter is to extend the research performed in Chapter 6 through the enhancement of the knowledge base used for performing the disambiguation (the co-occurrence graph) using multilingual information from parallel and comparable corpora. As we stated before, the biomedical WSD task is monolingual, hence in this case multilinguality can be seen as additional evidence which can be used for performing the disambiguation. As we stated in that chapter, the co-occurrence graphs are built with both textual information and CUIs from the UMLS database, since the identifiers of biomedical concepts are only available in English in UMLS. By using nouns and adjectives from the abstracts selected for creating the graph, we are able to include multilingual information in one or more additional languages into the graph. Results obtained in the study described in the chapter indicate that there are cases in which multilinguality is able to provide a statistically significant improvement in the accuracy achieved by the CO-Graph system when tackling the biomedical WSD task. That improvement is higher when a small subset of documents from the corpus is used for creating the co-occurrence graph, while, for graphs created with a greater number of documents, the improvements are reduced until the accuracy achieved is somehow similar to that obtained with monolingual information only. One of the main conclusions that can be extracted from this fact is that the improvement provided by the inclusion of multilingual data to the knowledge base used for performing WSD in the biomedical domain, is similar to the improvement that can be achieved by adding more information in a monolingual way. That is, a monolingual system would need a significantly larger amount of information than the multilingual one for achieving similar results. From this conclusion we can then derive one the most important findings of this thesis: **the proposed CO-Graph technique enhanced with multilingual information is able to offer better results than monolingual systems for biomedical WSD tasks, in scenarios in which the amount of available information (represented in this case by the number of documents composing the corpus) is reduced.**

Research Question 9: *Is multilingual data usually available in the biomedical domain? If not, which is the best way to automatically supply multilingual information to a system performing WSD in this domain?*

In the study presented in Chapter 7 we can also find enough evidence for answering this last Research Question. We observe that multilingual data, and particularly parallel and comparable corpora manually written and translated into different languages, are really difficult to find in the literature. This lack of multilingual resources have forced us to develop our own bilingual comparable corpora, written in both English and Spanish, and generated from abstracts downloaded from Ibero-American Elsevier journals which ask authors to write the abstracts of their papers in English and Spanish for publishing them. Motivated by this lack of multilingual information, we studied the possibility of automatically translating biomedical corpora originally written in English, and test our system with these automatic

translations. For this purpose, we used the Yandex translator over the NLM-related corpus (previously used in Chapter 6 for building the co-occurrence graph), with a range of different languages presenting different roots and characteristics. The achieved results proved that, for the biomedical WSD task, automatically created multilingual comparable corpora are able to achieve similar results to those obtained by manually created multilingual corpora. This way, automatic translation arises as a very useful way of generating multilingual information for this kind of tasks.

8.3 Future Lines of Work

Many new questions have arised during the development of all the different parts of this research. In this final section, some ideas about possible extensions and future steps regarding those new lines of work will be briefly mentioned.

We have obtained some interesting conclusions, already described in Section 8.2, regarding the disambiguation algorithms that have been used in the different tasks of this thesis. In particular, we have proved that those algorithms involving weights of the graph usually perform better than techniques that only take the structural information of the graph into account. However, as we have also stated in previous chapters, the differences between those algorithms have been always small. More work can be done in refining those algorithms used for combining the information offered by the co-occurrence graph and the context of each new instance, both in the CLWSD task and in the biomedical WSD task. Also, different disambiguation algorithms, apart from those already studied, can offer new possibilities to the improvement of the proposed technique. For example, combining algorithms mainly focused on structural information such as PageRank and methods relying on quantitative values (weights of the links, for example) such as Dijkstra's could be a good starting point for this line of research. Following this idea, community-based algorithms, which have offered successful results in the CLWSD task, should also be tested in the biomedical WSD task as well.

Regarding the Cross-Lingual Word Sense Disambiguation task, considering translations containing more than one word (multi-word translations) is a necessary step for next versions of the CO-Graph technique. More particularly, an analysis of the Gold-Standards used for evaluation indicates that in average, 3.26% of the expected translations are composed by more than one word. This value varies depending on the language (0.10% for German, 4.22% for Spanish, 1.75% for French, 8.50% for Italian and 2.36% for Dutch), so the potential improvement that can be achieved by incorporating these multi-word translations would also vary.

A deeper exploration of the dictionary extracted with GIZA++ is another line of work that should be explored in the near future, in order to include more possible translations that

would ideally allow the system to reach higher accuracy, as suggested in the analysis of performance of an ideal system presented in Section 5.5 of Chapter 5. A good approach could be to avoid forcing the number of maximum translations to be a fixed value, and instead varying that value depending on the statistical characteristics of the translations, or even selecting the value through an adaptative process.

Although we have already analysed the influence of selecting other types of words such as verbs, an exhaustive study of the impact of selecting different parts of speech for building the co-occurrence graph would possibly offer an insight on the importance of those words (adjectives, adverbs, etc.) for representing knowledge in a structured way.

Regarding multilinguality for biomedical WSD, future lines of work include the analysis of similar tasks when the original language is not English, but other languages that may present less available resources. Also, possible cross-lingual tasks for disambiguating a term written in a given language into its most suitable translation in a different target language should be explored. We also plan to apply the multilingual techniques described in this work to other tasks such as relation extraction, and in general to larger NLP systems performing more complex tasks which need WSD in their pipelines, for example, automatic text summarization.

Finally, the combination of our algorithm with other techniques used by similar state-of-the-art systems could also offer improvements in the different tasks addressed in this thesis.

8.4 Publications

The results of our research have been published in various conferences and international journals during the development of this thesis:

- Andres Duque, Juan Martinez-Romo, Lourdes Araujo. 2016. *Can Multilinguality Improve Biomedical Word Sense Disambiguation?*. Journal of Biomedical Informatics. In press. doi: http://dx.doi.org/10.1016/j.jbi.2016.10.020
- Juan Martinez-Romo, Lourdes Araujo, Andres Duque. 2016. *SemGraph: Extracting keyphrases following a novel semantic graph-based approach*. Journal of the Association for Information Science and Technology, 67(1), pp. 71-82.
- Andres Duque, Juan Martinez-Romo, Lourdes Araujo. 2015. *Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation*. Knowledge-Based Systems, 81, pp. 65-76.
- Andres Duque, Lourdes Araujo, Juan Martinez-Romo. 2015. *CO-Graph: A new graph-based technique for cross-lingual word sense disambiguation*. Natural Language Engineering, 21(5), pp. 743-772.

- Andres Duque, Juan Martinez-Romo, Lourdes Araujo. 2015. *Extracción no supervisada de relaciones entre medicamentos y efectos adversos*. Procesamiento del Lenguaje Natural, 55, pp. 83-90.
- Andres Duque, Lourdes Araujo, Juan Martinez-Romo. 2013. *Una nueva técnica de construcción de grafos semánticos para la desambiguación bilingüe del sentido de las palabras*. Procesamiento del Lenguaje Natural, 51, pp. 187-194.

BIBLIOGRAPHY

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2013. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics* 40.

Eneko Agirre and Philip Edmonds. 2007a. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Eneko Agirre and Philip Glenny Edmonds. 2007b. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Eneko Agirre and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7*. Association for Computational Linguistics, page 3.

Eneko Agirre, David Martínez, Oier López De Lacalle, and Aitor Soroa. 2006a. Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm. In *Proceedings of the first workshop on graph based methods for natural language processing*. Association for Computational Linguistics, pages 89–96.

Eneko Agirre, David Martínez, Oier Lopez de Lacalle, and Aitor Soroa. 2006b. Two graph-based algorithms for state-of-the-art wsd. In *EMNLP*. pages 585–593.

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 16–22.

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pages 7–12.

Eneko Agirre and Aitor Soroa. 2008. Using the multilingual central repository for graph-based word sense disambiguation. In *LREC*.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '09, pages 33–41.

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics* 26(22):2889–2896.

Marianna Apidianaki. 2008. Translation-oriented word sense induction based on parallel corpora. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual wsd and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '09, pages 77–85.

Marianna Apidianaki. 2013. Limsi: Cross-lingual word sense disambiguation using translation sense clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. *Proceedings of the American Medical Informatics Association (AMIA)* pages 17–21.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25(1):25–29.

Jordi Atserias, Luıs Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository .

Ségolène Aymé. 2016. The importance of review articles in making the voice of rare diseases heard: Ojrd's 10th anniversary. *Orphanet Journal of Rare Diseases* 11(1):1–2.

Carmen Banea and Rada Mihalcea. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics*. Association for Computational Linguistics, IWCS '11, pages 25–34.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai*. volume 3, pages 805–810.

Regina Barzilay and Michael Elhadad. 1999. Using lexical chains for text summarization. *Advances in automatic text summarization* pages 111–121.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *journal of machine learning research* 3(Feb):1137–1155.

Chris Biemann. 2006. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, TextGraphs-1, pages 73–80.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* 32(suppl 1):D267–D270.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, pages 1247–1250.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, pages 144–152.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *COMPUTER NETWORKS AND ISDN SYSTEMS*. Elsevier Science Publishers B. V., pages 107–117.

Marine Carpuat. 2013. Nrc: A machine translation approach to cross-lingual word sense disambiguation (semeval-2013 task 10). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 188–192.

Joyce Yue Chai and Alan W Biermann. 1999. The use of word sense disambiguation in an information extraction system. In *AAAI/IAAI*. pages 850–855.

Rachel Chasin, Anna Rumshisky, Ozlem Uzuner, and Peter Szolovits. 2014. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. *Journal of the American Medical Informatics Association* 21(5):842–849.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, pages 160–167.

Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, pages 133–140.

Bharath Dandala, Rada Mihalcea, and Razvan Bunescu. 2013a. Multilingual word sense disambiguation using wikipedia. Asian Federation of Natural Language Processing, Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 498–506.

Bharath Dandala, Rada Mihalcea, and Razvan Bunescu. 2013b. Word sense disambiguation using wikipedia. In *The People's Web Meets NLP*, Springer, pages 241–262.

Mona T. Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL*. pages 255–262.

E. W. Dijkstra. 1959. A note on two problems in connexion with graphs. *NUMERISCHE MATHEMATIK* 1(1):269–271.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS* 19(1):61–74.

Robert M Fano and David Hawkins. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics* 29(11):793–794.

Manaal Faruqui. 2014. " translation can't change a name": Using multilingual data for named entity recognition. *arXiv preprint arXiv:1405.0701* .

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Erwin Fernandez-Ordonez, Rada Mihalcea, and Samer Hassan. 2012. Unsupervised word sense disambiguation with multilingual representations. In *LREC*. pages 847–851.

Ronald Aylmer Fisher. 1925. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Ronald Aylmer Fisher. 1929. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 125(796):54–59.

William A Gale, Kenneth W Church, and David Yarowsky. 1992a. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26(5-6):415–439.

William A Gale, Kenneth W Church, and David Yarowsky. 1992b. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, pages 233–237.

Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *IJCAI*. volume 3, pages 1486–1488.

Sylvain Gaudan, Harald Kirsch, and Dietrich Rebholz-Schuhmann. 2005. Resolving abbreviations to their senses in medline. *Bioinformatics* 21(18):3658–3664.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*. pages 2525–2529.

Weiwei Guo and Mona Diab. 2010. Coleur and colslm: A wsd approach to multilingual lexical substitution, tasks 2 and 3 semeval 2010. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '10, pages 129–133.

Weiwei Guo and Mona T Diab. 2009. Improvements to monolingual english word sense disambiguation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics, pages 64–69.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. 2005. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* 33(suppl 1):D514–D517.

Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '02, pages 517–526.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database* 305:305–332.

David B. Hitchcock. 2009. Yates and contingency tables: 75 years later. *Journal Électronique d'Histoire des Probabilités et de la Statistique [electronic only]* 5(2):1–14; electronic only.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

David Hope and Bill Keller. 2013. Maxmax: a graph-based soft clustering algorithm applied to word sense induction. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 368–381.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pages 7304–7308.

Susanne M. Humphrey, Willie J. Rogers, Halil Kilicoglu, Dina Demner-fushman, and Thomas C. Rindflesch. 2006. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. *J. Am. Soc. Inform. Sci. Tech* 57:96–113.

Betsy L Humphreys, Donald AB Lindberg, Harold M Schoolman, and G Octo Barnett. 1998. The unified medical language system. *Journal of the American Medical Informatics Association* 5(1):1–11.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. ACL.

Nancy Ide and Jean Veronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics* 24:1–40.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008* .

Antonio Jimeno-Yepes and Alan R. Aronson. 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics* 11:569.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics* 12(1):223.

Mahesh Joshi, Serguei Pakhomov, Ted Pedersen, and Christopher G Chute. 2006. A comparative study of supervised learning as applied to acronym expansion in clinical reports. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2006, page 399.

Dimitar Kazakov and R. Ahmad Shahid. 2013. Using parallel corpora for word sense disambiguation. INCOMA Ltd. Shoumen, BULGARIA, Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pages 336–341.

Dimitar Lubomirov Kazakov and Ahmad Raza Shahid. 2010. Retrieving lexical semantics from multilingual corpora. *Polibits* pages 25–28.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities* 34(1-2):15–48.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology* 105:116.

Ioannis P. Klapaftis and Suresh Manandhar. 2008. Word sense induction using graphs of collocations. In *Proceeding of the 2008 conference on ECAI 2008*. IOS Press, Amsterdam, The Netherlands, The Netherlands, pages 298–302.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. volume 5.

Claudia Leacock, George A Miller, and Martin Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* 24(1):147–165.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pages 41–48.

Els Lefever and Veronique Hoste. 2010a. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '10, pages 15–20.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 158–166.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 317–322.

Els Lefever and Véronique Hoste. 2010b. Construction of a benchmark data set for cross-lingual word sense disambiguation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*. ACM, pages 24–26.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 768–774.

Hongfang Liu, Yves A. Lussier, and Carol Friedman. 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method. *Journal of Biomedical Informatics* 34(4):249 – 261.

Fernando López-Ostenero. 2002. *Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario*. Ph.D. thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia.

Suresh Manandhar, Ioannis P Klapaftis, Dmitriy Dligach, and Sameer S Pradhan. 2010. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics, pages 63–68.

Lluís Màrquez, Gerard Exsudero, David Martínez, and German Rigau. 2006. Supervised corpus-based methods for wsd. In *Word Sense Disambiguation: Algorithms and Applications*, Springer, Dordrecht, The Netherlands, volume 33 of *Text, Speech and Language Technology*, pages 167–216.

David Martínez, Eneko Agirre, and Lluís Màrquez. 2002. Syntactic features for high precision word sense disambiguation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 1–7.

Juan Martinez-Romo, Lourdes Araujo, Javier Borge-Holthoefer, Alex Arenas, José A. Capitán, and José A. Cuesta. 2011. Disentangling categorical relationships through a graph of co-occurrences. *Phys. Rev. E* 84:046108.

Margaret Masterman. 1961. Semantic message detection for machine translation, using an interlingua. In *Proc. 1961 International Conf. on Machine Translation*. pages 438–475.

Alexa T McCray, Suresh Srinivasan, and Allen C Browne. 1994. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, page 235.

Bridget T McInnes. 2008. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*. Association for Computational Linguistics, pages 49–54.

Bridget T McInnes, Ted Pedersen, Ying Liu, Serguei V Pakhomov, and Genevieve B Melton. 2011. Using second-order vectors in a knowledge-based method for acronym disambiguation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 145–153.

Bridget T McInnes and Mark Stevenson. 2014. Determining the difficulty of word sense disambiguation. *Journal of biomedical informatics* 47:83–90.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '05, pages 411–418.

Rada Mihalcea. 2006. Knowledge-based methods for wsd. In *Word Sense Disambiguation: Algorithms and Applications*, Springer, Dordrecht, The Netherlands, volume 33 of *Text, Speech and Language Technology*, pages 107–132.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 1126.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pages 303–308.

Sungrim Moon, Bjoern-Toby Berster, Hua Xu, and Trevor Cohen. 2013. Word sense disambiguation of clinical abbreviations with hyperdimensional computing. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2013, page 1007.

Sungrim Moon, Serguei Pakhomov, and Genevieve B Melton. 2012. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2012, page 1310.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics* 17(1):21–48.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41(2):10.

Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(4):678–692.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '10, pages 216–225.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP-CoNLL '12, pages 1399–1410.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654* .

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 40–47.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.* 29(1):19–51.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 613–619.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50(302):157–175.

Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. 2005. Maximizing semantic relatedness to perform word sense disambiguation. *University of Minnesota supercomputing institute research report UMSI* 25:2005.

Andrew Philpot, Eduard Hovy, and Patrick Pantel. 2005. The omega ontology. In *Proceedings, IJCNLP workshop on Ontologies and Lexical Resources (OntoLex-05)*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Int'l Conference on Global WordNet*. Citeseer.

Laura Plaza, Mark Stevenson, and Alberto Díaz. 2012. Resolving ambiguity in biomedical text to improve summarization. *Information Processing & Management* 48(4):755–766.

P. Pons and M. Latapy. 2005. Computing communities in large networks using random walks. *Lect. Notes Comput. Sci.* 3733:284.

Judita Preiss. 2006. *Probabilistic word sense disambiguation: analysis and techniques for combining knowledge sources*. Ph.D. thesis, University of Cambridge.

Judita Preiss and Mark Stevenson. 2015. The effect of word sense disambiguation accuracy on literature based discovery. In *Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics*. ACM, pages 1–1.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1(1):81–106.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 109–117.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* .

Philip Resnik. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In *Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing)*. pages 283–299.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering* 5(2):113–133.

Ronald L Rivest. 1987. Learning decision lists. *Machine learning* 2(3):229–246.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases....* TY Crowell Company.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127* .

Alex Rudnick, Can Liu, and Michael Gasser. 2013. Hltdi: Cl-wsd using markov random fields for semeval-2013 task 10. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 171–177.

John-Arne Røttingen, Sadie Regmi, Mari Eide, Alison J Young, Roderik F Viergever, Christine Årdal, Javier Guzman, Danny Edwards, Stephen A Matlin, and Robert F Terry. 2013. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *The Lancet* 382(9900):1286 – 1307.

Guergana K. Savova, Anni R. Coden, Igor L. Sominsky, Rie Johnson, Philip V. Ogren, Piet C. de Groen, and Christopher G. Chute. 2008. Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics* 41(6):1088 – 1100.

Eric Sayers. 2013. A general introduction to the e-utilities .

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*. Manchester, UK, volume 12, pages 44–49.

Martijn J Schuemie, Jan A Kors, and Barend Mons. 2005. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology* 12(5):554–565.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics* 24(1):97–123.

Aditya K Sehgal, Padmini Srinivasan, and Olivier Bodenreider. 2004. Gene terms and english words: An ambiguous mix. In *Proc. of the ACM SIGIR Workshop on Search and Discovery for Bioinformatics, Sheffield, UK*. Citeseer.

Carina Silberer and Simone Paolo Ponzetto. 2010. Uhd: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '10, pages 134–137.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, page 662.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*. pages 2142–2147.

Angus Stevenson. 2010. *Oxford dictionary of English*. Oxford University Press, USA.

Mark Stevenson and Yikun Guo. 2010a. Disambiguation in the biomedical domain: The role of ambiguity type. *Journal of Biomedical Informatics* 43(6):972 – 981.

Mark Stevenson and Yikun Guo. 2010b. Disambiguation of ambiguous biomedical terms using examples generated from the umls metathesaurus. *Journal of biomedical informatics* 43(5):762–773.

Mark Stevenson, Yikun Guo, Abdulaziz Al Amri, and Robert Gaizauskas. 2009. Disambiguation of biomedical abbreviations. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, BioNLP '09, pages 71–79.

Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. *Proceedings of NAACL HLT 2015* pages 314–323.

Liling Tan and Francis Bond. 2013. Xling: Matching query sentences to a parallel corpus using topic models for wsd. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 167–170.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*. volume 5, pages 237–248.

Stéphan Tulkens, Simon Šuster, and Walter Daelemans. 2016. Using distributed representations to disambiguate biomedical and clinical concepts. *arXiv preprint arXiv:1608.05605* .

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pages 384–394.

Carlos Valmaseda, Juan Martinez-Romo, and Lourdes Araujo. 2016. A tagged corpus for automatic labeling of disabilities in medical scientific papers. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Maarten van Gompel. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '10, pages 238–241.

Maarten van Gompel and Antal van den Bosch. 2013. Wsd2: Parameter optimisation for memory-based cross-lingual word-sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 183–187.

Jean Veronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language* 18(3):223–252.

Darnes Vilariño, Carlos Balderas, David Pinto, Miguel Rodríguez, and Saul León. 2010. Fcc: Modeling probabilities with giza++ for task #2 and #3 of semeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '10, pages 112–116.

Piek Vossen. 1996. Right or wrong. combining lexical resources in the eurowordnet project. In *M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, CR Papmehl, Proceedings of Euralex-96, Goetheborg*. Citeseer, pages 715–728.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Thomas Wasow, Amy Perfors, and David Beaver. 2005. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe* pages 265–282.

Warren Weaver. 1955. Translation. *Machine translation of languages* 14:15–23.

Marc Weeber, James G. Mork, and Alan R. Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA 2001 Symposium*. pages 746–750.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 1–7.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1(6):80–83.

Yorick Wilks. 1968. *On-line semantic analysis of English texts*. Citeseer.

Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical abbreviation disambiguation using neural word embeddings. *ACL-IJCNLP 2015* page 171.

Hua Xu, Peter D Stetson, and Carol Friedman. 2012. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2012, page 1004.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pages 266–271.

Dayu Yuan, Ryan Doherty, Julian Richardson, Colin Evans, and Eric Altendorf. 2016. Word sense disambiguation with neural language models. *arXiv preprint arXiv:1603.07012* .

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, pages 78–83.